Resource Allocation in NOMA-based Self-Organizing Networks using Stochastic Multi-Armed Bandits

Marie-Joséph a Youssef, Venugopal V. Veeravalli, Joumana Farah, Charbel Abdel Nour, Catherine Douillard

Resource Allocation in NOMA-based Self-Organizing Networks using Stochastic Multi-Armed Bandits

Marie-Josepha Youssef, *Member, IEEE,* Venugopal V. Veeravalli, *Fellow, IEEE,* Joumana Farah, *Member, IEEE,* Charbel Abdel Nour, *Senior Member, IEEE,* and Catherine Douillard, *Senior Member, IEEE*

***Abstract*—To achieve better connectivity in future communication networks, the deployment of different types of access points (APs) is underway. APs are expected to be equipped with self-organizing capabilities to reduce costs. Moreover, due to the spectrum crunch, frequency reuse among the deployed APs is inevitable, exacerbating the problem of inter-cell interference (ICI). Therefore, ICI mitigation in self-organizing networks (SONs) is commonly identified as a key radio resource management mechanism to enhance performance. To this end, this paper proposes a novel solution for the uncoordinated channel and power allocation problems. Based on the multi-armed bandits (MAB) framework, the proposed technique does not require any communication between the APs. The case of varying channel rewards across APs is considered. In contrast to previous work on channel allocation using the MAB framework, APs are permitted to choose multiple channels for transmission. Moreover, non-orthogonal multiple access is used, allowing multiple APs to access each channel simultaneously. This results in an MAB model with varying channel rewards, multiple plays and non-zero reward on collision. The proposed algorithm has an expected regret in the order of $\mathcal{O}(\log^2 T)$, with extensive numerical results revealing it significantly outperforms a well-known baseline algorithm in terms of energy efficiency.**

***Index Terms*—Uncoordinated channel and power allocation, MAB with multiple plays and non-zero reward on collision, varying reward distribution, NOMA, self-organizing networks.**

I. Introduction

Future cellular communication networks are expected to support a myriad of new applications and services conceived for both traditional human-type devices and for the growing number of machine-type devices [1]. To meet the exponential growth in connectivity and mobile traffic, new technologies are needed. Among these new technologies, the deployment of different types of access points (AP), e.g., small base-stations (SBS), pico-cells, femto-cells, relays, etc., is of particular importance, since APs can offload mobile traffic from highly congested macro-base stations (MBS) [2]. To limit human intervention and reduce planning and maintenance costs, APs can be equipped with self-organizing capabilities [3], allowing them to optimize their resource use in a distributed manner. APs normally have a lower transmit power budget and a smaller coverage range when compared to traditional MBSs. However, thanks to their denser deployment, APs benefit from the ability to consume less transmit power, leading to significant gains in power consumption as was shown in [4], [5]. That said, by introducing APs into the network, the problem of inter-cell-interference (ICI) is aggravated, necessitating the application of adequate resource allocation algorithms to limit the interference [6].

The problem of ICI in self-organizing networks (SON) was extensively studied in the literature. In [7], the weighted sum-rate of the system is optimized through ICI coordination between SBSs. The authors adopt a blanking method where at the level of each SBS, some wireless channels are not used to mitigate the ICI. In [8], an algorithm for ICI coordination between SBSs based on asynchronous inter-cell signaling is proposed. The authors of [9] propose an algorithm based on a semi-static frequency allocation to mitigate ICI and enhance the performance of cell-edge users. The proposed solutions of [7]–[9] rely on explicit communication between the distributed SBSs to mitigate ICI, resulting in excessive signaling among SBSs. To limit signaling overhead, decentralized algorithms, based on reinforcement learning, are preferred.

The use of reinforcement learning in wireless communications has recently garnered significant attention [10]. The related framework of multi-player multi-armed bandits (MAB) [11] has also been widely used to study multiple problems in wireless communication systems ranging from SON [12]–[14], to uncoordinated spectrum access [15]–[18], to fast uplink grant allocation [19], to unmanned-aerial vehicles positioning and path-planning [20]. In the context of SON, in [12], [13], a solution is proposed based on the stochastic MAB framework to allow SBSs to partition efficiently the available frequency resources in an effort to mitigate ICI. In [21], a method based on learning automata is proposed where femto-cells adjust their resource use based on the feedback received from users. In [14], the authors resort to the EXP3 algorithm from the adversarial MAB framework to mitigate the ICI while allowing each base-station (BS) to access multiple frequency bands. The work in [22] proposes a data-driven approach based on the MAB framework to address the ICI problem in heterogeneous networks (HetNets). The MAB framework was also widely used to study the opportunistic

M. J. Youssef, C. Abdel Nour and C. Douillard are with IMT Atlantique, LabSTICC, UBL, F-29238 Brest, France, (e-mail: marie-josepha.youssef@imt-atlantique.fr; charbel.abdelnour@imt-atlantique.fr; catherine.douillard@imt-atlantique.fr).

V. V. Veeravalli is with the ECE Department, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA, and also with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: vvv@illinois.edu).

J. Farah is with the Department of Electricity and Electronics, Faculty of Engineering, Lebanese University, Roumieh, Lebanon (e-mail: joumana-farah@ul.edu.lb).

This work has been funded with support from the UBL, the GdR ISIS, the Lebanese University, and the US National Science Foundation SpecEES program under grant number 1730882, throughout the University of Illinois at Urbana-Champaign (UIUC).

and the uncoordinated spectrum access problems. For example, in [15], [16] and [23], the MAB model is used to study the opportunistic spectrum access problem in cognitive radio networks where secondary users compete to access the part of the spectrum not occupied by primary users. In addition to studying the opportunistic channel access problem, in [23], the authors also solve the distributed power allocation problem. In contrast to opportunistic channel access, the authors of [17], [18] and [24] employ MAB to study the uncoordinated spectrum access problem without distinguishing between the users. The distributed power control problem is studied in [24], and solutions are proposed based on the upper-confidence-bound (UCB) algorithm, and on the ϵ-greedy algorithm. In [25], the channel and power allocation problem in a device-to-device system is modeled using the MAB framework. A game-theoretic solution based on the potential game framework is proposed to minimize regret of users.

With the exception of [14], all previous work on wireless communications solutions based on MABs assumes that each player chooses one channel at each timeslot. However, removing this assumption is expected to improve performance for the players if a suitable algorithm is formulated, especially for the case of a SON. Indeed when an AP can access multiple channels simultaneously, an increase in both, the probability of a successful transmission and the achieved reward or rate is observed, allowing the AP to serve more end-users. Moreover, with the exception of [17], [18], [25], all previous work based on MABs considered a zero reward for multiple players accessing the same channel. By alleviating this assumption and adopting non-orthogonal multiple access (NOMA), system performance is expected to further improve.

From an information-theoretical point of view, it is well-known that non-orthogonal user multiplexing using superposition coding at the transmitter and proper decoding techniques at the receiver not only outperforms orthogonal multiplexing, but is also optimal in the sense of achieving the capacity region of the downlink broadcast channel [26]. As a result, NOMA emerged as a promising multiple access technology for 5G systems [27]–[29]. NOMA allows multiple users to be scheduled on the same time-frequency resource by multiplexing them in the power domain. At the receiver side, successive interference cancellation (SIC) is performed to retrieve superimposed signals.

To limit the ICI in a SON, studying the resource allocation in the fronthaul portion of the network is of utmost importance [12]–[14]. When coupled with optimizing the resource allocation in the backhaul link, optimizing the fronthaul portion leads to significant performance gains [29], [30].

In this paper, we consider the fronthaul part of a self-organizing wireless network where multiple APs aim at organizing their uplink transmissions with a central unit in a distributed manner. Both the uncoordinated channel access and the distributed power control problems are studied. A solution based on the MAB framework, which does not necessitate any coordination or communication between APs, is proposed. The considered setting is closest to the ones studied in [17] and [31], where a game-theoretic approach is used to solve the uncoordinated channel access problem. Our study extends that of [17] and [31] by allowing each AP to access multiple channels simultaneously and by proposing a model for the distributed power control problem. The main contributions of this paper can be summarized as follows:

- A two-phase algorithm based on the MAB framework, extending the work in [17], [31], is proposed for the uncoordinated channel access and distributed power control problems.
- For the first phase, i.e., the uncoordinated channel access phase, in addition to considering varying channel rewards between APs, each AP is allowed to simultaneously access multiple channels. This is in contrast to the work in [17] and [31] where each player accesses one channel in a timeslot. Moreover, each channel can accommodate multiple APs at once using NOMA, leading to a multi-player MAB problem with varying player rewards, multiple plays and non-zero reward on collision.
- For the power control phase, varying power level rewards between APs are considered and an algorithm to solve the power control problem on each channel is proposed.
- The proposed technique is shown to achieve a sublinear regret of $\mathcal{O}(\log^2 T)$. In addition, simulation results validating the theoretical results and the performance of the proposed technique are presented.
- To the best of our knowledge, this is the first work that studies the uncoordinated channel access and the distributed power control problems in a SON network, using both NOMA and the multi-player MAB framework with varying channel rewards across users, multiple plays, and non-zero reward on collision.

The rest of this paper is organized as follows. The system model is presented in section II. In sections III, IV, V and VI, the proposed algorithm is presented along with an analysis of the system-wide regret. Simulation results are presented in section VII and conclusions in section VIII.

II. System Model

Consider the uplink of a cellular system as shown in Fig. 1 where K APs aim to organize their communications with an MBS serving as gateway to the core network, over M available wireless channels, in an uncoordinated manner. The communication occurs over a finite time horizon T that may not be known in advance to the APs. At each timeslot t, every AP k chooses N channels, adjusts its transmission power, and transmits over the chosen channels. Note that the proposed solution can be easily extended to the case where each AP $k \in \mathcal{K}$ chooses N_k channels at each timeslot, where $1 \leq N_k \leq M$. We assume that NOMA is employed, enabling multiple APs to choose the same channel for communication and achieve a non-zero rate. That said, if two or more APs choose the same channel, the received power levels of these APs must be different at the receiving BS level in the core network, to enable SIC decoding at the receiver side. To ensure the reception of different received power levels for the signals transmitted by the APs, we generalize the uplink NOMA power allocation model introduced in [32], where for a constant SINR requirement, L received power levels, ensuring

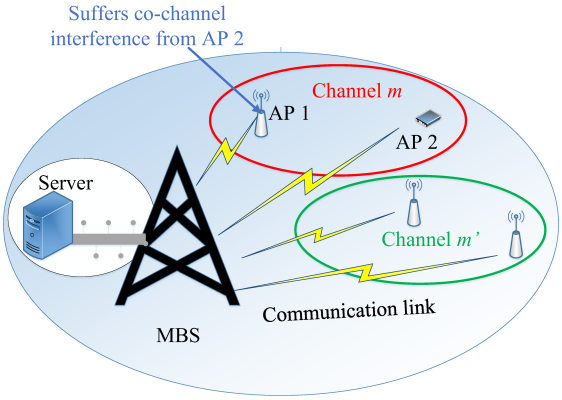


Fig. 1: System Model.

the SINR requirement for L users scheduled on the same channel, are calculated. In this work, we extend the study of [32] to allow for L distinct SINR requirements per channel, $\mathbf{\Gamma} = \{\Gamma_1, \ldots, \Gamma_L\}$, sorted by decreasing order. Note that allowing for distinct SINR levels inherently encompasses the special case of constant SINR levels.

An AP k choosing SINR requirement Γ_l over channel m achieves the following uplink data rate:

$$R_{k,m,l} = \log_2(1 + \Gamma_l), \tag{1}$$

where Γ_l is given by:

$$\Gamma_l = \frac{v_l}{V_l + N_0 B_c}. \tag{2}$$

In Eq. (2), v_l is the received power level of AP k, the expression of which is given in Section II-B, N_0 is the noise power spectral density and B_c the channel bandwidth. At the receiver side, when the AP transmissions are received with different power levels, SIC is employed to decode the received messages in a descending order. In other words, the AP choosing a higher SINR requirement Γ_l, and consequently a higher received power level v_l, suffers interference from all APs choosing a lower SINR requirement. Once decoded, the signal of the AP choosing Γ_l is removed using SIC before decoding the remaining messages. Hence, variable V_l of Eq. (2) is the power level of the interfering transmissions, not canceled with SIC, expressed as: $V_l = \sum_{l'=l+1}^{L} v_{l'}$. To limit the decoding complexity at the receiving BS in the core network, as well as the error propagation in SIC, the number of APs allowed to access a channel and achieve a non-zero rate is limited to β, such that $\beta M \geq KN$. Note that in the case of a varying number of chosen channels across users, this last condition becomes $\beta M \geq \sum_{k \in \mathcal{K}} N_k$. It is assumed that when an AP k accesses a channel m at timeslot t, k receives feedback at the end of the timeslot regarding the total number of APs currently accessing channel m. No *a priori* knowledge of the channel gain experienced over each channel is assumed. Moreover, these channel gains are distinct for each AP. To solve the channel and power allocation problems in an uncoordinated manner, we proceed in two steps, the first, of length T_C, dedicated to channel allocation and the second, of length T_P, dedicated to power allocation. Note that both T_C and T_P may not be known to the APs.

A. Uncoordinated Channel Allocation

To allow each AP to access N channels simultaneously in a NOMA manner, the problem of uncoordinated multiple access is modeled as a stochastic multi-player MAB problem with multiple plays and non-zero reward on collision. The set of players is the set of APs $\mathcal{K}$ and the set of arms is the set of channels $\mathcal{M}$. The action of each AP k at each timeslot t is $\boldsymbol{a}_k^t \in \{0,1\}^{M \times 1}$ such that $a_k^t(m) = 1$ if AP k pulls channel m at timeslot t. Moreover, $\sum_{m=1}^{M} a_k^t(m) = N$, $\forall k \in \mathcal{K}$. The action space of each AP k, $\mathcal{A}_k$, consists of all possible combinations of N channels, hence $|\mathcal{A}_k| = \binom{M}{N}$. Let $\boldsymbol{a}^t = \{\boldsymbol{a}_1^t, \ldots, \boldsymbol{a}_K^t\}$ denote the strategy profile of all APs in timeslot t. Upon choosing an action $\boldsymbol{a}_k^t \in \boldsymbol{a}^t$, AP k receives the following average reward:

$$g_k^t(\boldsymbol{a}^t) = \sum_{m=1}^{M} a_k^t(m) \mu_M(k, m, k_m), \tag{3}$$

where k_m is the number of APs choosing channel m at timeslot t. The variable $\mu_M(k, m, k_m)$ is the mean reward of AP k over channel m when k_m APs access it. Note that the actual value of the received reward by AP k when choosing channel m at timeslot t is drawn from a uniform distribution with mean $\mu_M(k, m, k_m)$.

We assume that the mean reward of AP k when accessing channel m alone is equal to the normalized average channel gain of AP k over channel m, i.e.,

$$\mu_M(k, m, 1) = h_{k,m} / \mu_M^{max}, \tag{4}$$

where $\mu_M^{max} = \max_{k \in \mathcal{K}, m \in \mathcal{M}} h_{k,m}$ and $h_{k,m}$ is the average channel gain of AP k over channel m accounting for both small scale Rayleigh fading and large scale fading (i.e., path-loss and log-normal shadowing). Note that it is assumed that the BS at the core network performs channel estimation on the received signals from all APs. Hence, the average channel gains $h_{k,m}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}$ are assumed to be perfectly known by the receiving BS. For $1 < k_m \leq \beta$, the mean reward of an AP must account for the added interference brought by the $(k_m - 1)$ other APs scheduled on the same channel m. Ideally, the mean reward should take into account the interference brought by each particular AP. However, that would result in a prohibitive complexity since any channel, for each $1 < k_m \leq \beta$, would have $\binom{K-1}{k_m}$ distinct reward values. To simplify the analysis, in this work, we assume that the mean reward for $1 < k_m \leq \beta$, is a decreasing function of the number of interfering APs on the same channel. In other words,

$$\mu_M(k, m, k_m) = \mu_M(k, m, 1) / k_m. \tag{5}$$

When $k_m > \beta$, $\mu_M(k, m, k_m) = 0$. The normalization in Eq. (4) leads to: $\mu_M(k, m, k_m) \in [0, 1]$ for every AP $k \in \mathcal{K}$, on every channel $m \in \mathcal{M}$ and for every number of APs $k_m \in [\beta]$. Hence, $g_k^t(\boldsymbol{a}^t) \in [0, N]$.

In addition to receiving the achieved rewards, we assume that the feedback received by each AP k from the MBS includes the total number of APs simultaneously accessing its chosen channels. In other words, for all channels m such that $a_k^t(m) = 1$, AP k receives the total number of APs

accessing channel m, i.e., receives $k_m = \sum_{k\in\mathcal{K}} a_k^t(m)$. This information is useful for the future decisions of APs regarding chosen channels and is necessary for the correct estimation of the mean rewards, allowing APs to learn and settle on the optimal allocation. Moreover, since β is normally kept small, feeding back to each AP k the total number of APs simultaneously accessing its chosen channels requires only a few bits.

APs make their decisions in a distributed manner observing neither the channels chosen by other APs nor the rewards received by other APs. Each AP k can only observe the reward it gets on each of its chosen channels. Our aim is to propose a distributed algorithm allowing APs to organize their transmissions on the available channels, without communicating together, in such a way as to maximize the sum reward of the system. By definition, the action profile yielding the highest sum reward $\boldsymbol{a}^*$ is given by:

$$\boldsymbol{a}^* = \underset{\boldsymbol{a}\in\mathcal{A}}{\operatorname{argmax}} \sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)\mu_M(k,m,k_m), \tag{6}$$

where $\mathcal{A}$ is the action space of all APs, i.e., $\mathcal{A} = \prod_{k\in\mathcal{K}} \mathcal{A}_k$.

The expected regret incurred during T_C is the difference between the achieved reward when playing $\boldsymbol{a}^*$ at all timeslots, and the actually achieved reward by the learning players during the T_C timeslots [11]. In our case, it is given by:

$$\begin{aligned}\bar{R} =& T_C \sum_{k,m} a_k^*(m)\mu_M(k,m,k_m^*) - \\ & \mathbb{E}\left(\sum_{t=1}^{T_C}\sum_{k,m} a_k(m)\mu_M(k,m,k_m)\right),\end{aligned} \tag{7}$$

where k_m^* is the optimal number of APs scheduled over channel m under $\boldsymbol{a}^*$.

After T_C timeslots, the APs receive a signal from the core network to terminate the channel allocation phase. At the end of the channel allocation phase, at most β APs are scheduled over each channel $m \in \mathcal{M}$. Moreover, as an outcome of this first phase, each AP k computes an estimate of its average channel gain over each channel m, denoted by $\hat{h}_{k,m}$.

B. Distributed Power Allocation

Once settled over their chosen channels, the APs receive a signal from the core network to move to the power allocation stage. Since different frequency bands are allocated to different channels, power allocation over each channel m can be done independently of other channels $m' \in \mathcal{M} \setminus \{m\}$. In the following, we will focus on the power allocation over channel $m \in \mathcal{M}$, where the set of scheduled APs is $\mathcal{K}_m$.

To simplify the distributed power allocation, we assume that each AP chooses, for each of its allocated channels, one SINR level among a fixed set of $L \geq \beta$ available SINR levels, with $\boldsymbol{\Gamma}$ being the set of pre-determined available SINR levels. The AP then calculates the necessary power level v_l for the chosen SINR level Γ_l. For successful SIC decoding, each power level can support one AP only. In other words, if multiple APs choose the same power level, SIC fails and the signals of all K_m APs are not decodable. Inspired by [32], it can be shown that, to satisfy Eq. (2), the power level v_l must be set as:

$$v_l = \Gamma_l N_0 B_c \prod_{l'=l+1}^{L} (\Gamma_{l'} + 1)\,. \tag{8}$$

Note that the expression of v_l is obtained by proceeding backwards and by induction from $v_L = \Gamma_L N_0 B_c$.

The expression of v_l ensures the SINR requirement Γ_l when considering that an AP chooses each subsequent SINR requirement, hence the worst case scenario. Note that our setting allows for similar SINR levels. However, for similar or distinct SINR levels, the power levels chosen by APs need to be distinct to allow for SIC decoding.

To ensure SIC stability, i.e., successful decoding of the received signals in descending order [33], the distributed power control scheme must ensure that the power of each signal scheduled for decoding at the BS is larger than the received power of the interference generated by the combination of the remaining signals, i.e., $v_l > V_l$. From Eq. (8), the power level v_l depends on the associated SINR level Γ_l as well as on the interfering SINR levels $\Gamma_{l'}, l' = l+1, \ldots, L$.

Proposition 1. To ensure SIC stability, the available SINR levels must satisfy:

$$\Gamma_l > \frac{2^{(L-l-1)} \times \Gamma_L}{\prod_{l'=l+1}^{L} (\Gamma_{l'} + 1)}. \tag{9}$$

Proof. By proceeding backwards, to get $v_{L-1} > v_L$, the following must hold:

$$\Gamma_{L-1} > \frac{\Gamma_L}{\Gamma_L + 1} = \frac{2^{(L-(L-1)-1)}\Gamma_L}{\Gamma_L + 1}. \tag{10}$$

Similarly, to get $v_{L-2} > v_{L-1} + v_L$, the following must hold:

$$\begin{aligned}\Gamma_{L-2} &> \frac{\Gamma_{L-1}(\Gamma_L+1)+\Gamma_L}{(\Gamma_{L-1}+1)(\Gamma_L+1)} \overset{\text{(a)}}{>} \frac{\frac{\Gamma_L}{\Gamma_L+1}(\Gamma_L+1)+\Gamma_L}{(\Gamma_{L-1}+1)(\Gamma_L+1)} \\ &> \frac{2\Gamma_L}{(\Gamma_{L-1}+1)(\Gamma_L+1)} = \frac{2^{(L-(L-2)-1)}\Gamma_L}{\prod_{l'=L-1}^{L}(\Gamma_{l'}+1)},\end{aligned} \tag{11}$$

where (a) follows from Eq. (10).

To get $v_l > V_l = \sum_{l'=l+1}^{L} v_{l'}$, assume that Eq. (9) holds. By induction, to get $v_{l-1} > \sum_{l'=l}^{L} v_{l'}$, we must have:

$$\Gamma_{l-1} > \frac{2^{(L-(l-1)-1)}\Gamma_L}{\prod_{l'=l}^{L}(\Gamma_{l'}+1)}. \tag{12}$$

■

Knowing the available SINR levels, each AP $k \in \mathcal{K}_m$ calculates the associated received power levels using Eq. (8). Then, using the estimated average channel gain over m, $\hat{h}_{k,m}$, AP $k \in \mathcal{K}_m$ calculates the necessary transmit power for each

power level v_l, $p_{k,m,l}$, according to:

$$p_{k,m,l} = v_l / \hat{h}^2_{k,m}. \tag{13}$$

Each AP is assumed to have a power budget per channel P^m_k. Hence, AP k can transmit over channel m using power level v_l if $p_{k,m,l} \leq P^m_k$. AP $k \in \mathcal{K}_m$ builds the set of possible power levels, $\mathcal{P}^a_{k,m}$, where $\mathcal{P}^a_{k,m} = \{v_l |\ p_{k,m,l} \leq P^m_k, l \in [L]\}$. Note that the set of possible power levels are AP-dependent because of their dependency on the estimated average channel gain of each AP, $\hat{h}_{k,m}$, and on the AP power budget.

The power allocation among APs on the same channel consists of APs choosing SINR levels, and hence received power levels, in a distributed manner, and without any inter-AP coordination. Since APs choosing the same SINR level result in an unsuccessful SIC decoding, the APs must aim at organizing their transmissions using different SINR levels. For this purpose, the power allocation on each channel is modeled using the MAB framework with single play and zero-reward on collision. Over channel m, the set of players is $\mathcal{K}_m$ and the set of arms is the set of power levels $\mathcal{VL} = \{v_l, l = 1, \ldots, L\}$. Since $L = |\mathcal{VL}| \geq \beta \geq K_m = |\mathcal{K}_m|$, a solution where each AP accesses one power level, without collision, is achievable. At each timeslot, each AP $k \in \mathcal{K}_m$ chooses an action $a^t_{k,m}$, i.e., a power level $v_l \in \mathcal{P}^a_{k,m}$, and transmits using $p_{k,m,l}$. The action space of AP k is $\mathcal{P}^a_{k,m}$. Let $\boldsymbol{a}^t_m$ denote the strategy chosen by all APs in $\mathcal{K}_m$ over channel m at timeslot t. Upon choosing action $a^t_{k,m} \in \boldsymbol{a}^t_m$, AP k receives the following average reward on channel m:

$$g^t_{k,m}(\boldsymbol{a}^t) = \mu_P(k, m, a^t_{k,m}) \eta(\boldsymbol{a}^t_m), \tag{14}$$

where $\mu_P(k, m, a^t_{k,m})$ is the reward of AP k when choosing $a^t_{k,m}$. Note that the actual value of the received reward by AP k when choosing action $a^t_{k,m}$ on channel m at timeslot t is drawn from a uniform distribution with mean $\mu_P(k, m, a^t_{k,m})$.

The mean reward $\mu_P(k, m, a^t_{k,m})$ is chosen in a way to strike a trade-off between SINR maximization and transmit power minimization. Therefore, it is set as:

$$\mu_P(k, m, a^t_{k,m} = v_l) = w^1_k \frac{\Gamma_l}{\Gamma_{max}} + w^2_k \frac{1}{p_{k,m,l}\ \max\limits_{k,m,l}(\frac{1}{p_{k,m,l}})}, \tag{15}$$

where w^1_k and w^2_k are weight parameters relative to AP $k \in \mathcal{K}_m$ satisfying $w^1_k + w^2_k = 1$. The variable Γ_{max} is the highest available SINR, i.e., $\Gamma_{max} = \Gamma_1$. Note that $\mu_P(k, m, a^t_{k,m}) \in [0, 1]$ and is not known by the AP in advance. Let $\mathcal{N}^m_{v_l}(\boldsymbol{a}^t_m)$ be the set of APs choosing power level v_l at timeslot t, i.e., $\mathcal{N}^m_{v_l}(\boldsymbol{a}^t_m) = \{k \in \mathcal{K}_m \mid a^t_{k,m} = v_l\}$. The variable $\eta(\boldsymbol{a}^t_m)$ is the collision indicator of the strategy profile of all APs, $\boldsymbol{a}^t_m$, i.e., $\eta(\boldsymbol{a}^t_m) = 1$ if $|\mathcal{N}^m_{a^t_{k,m}=v_l}(\boldsymbol{a}^t_m)| \leq 1, \forall\ v_l \in \mathcal{VL}$, and 0 otherwise. Note that no feedback regarding the value of the collision indicator $\eta(\boldsymbol{a}^t_m)$ is necessary. In fact, in the case of collision on channel m, the MBS returns a zero reward to the APs having chosen channel m. When no collision takes place, the MBS returns only the value of the mean reward to the AP since the collision indicator is equal to one in the case of no collision.

APs choose power levels in a distributed manner without any coordination, with each AP only observing the reward received on the chosen power level. The proposed power allocation scheme aims at maximizing the sum reward of the system. Let $\boldsymbol{a}^{*P}_m$ be the action profile yielding the highest sum reward over channel m:

$$\boldsymbol{a}^{*P}_m = \underset{\boldsymbol{a}_m \in \mathcal{P}^a_m}{\operatorname{argmax}} \sum_{k \in \mathcal{K}_m} \mu_P(k, m, a^t_{k,m})\, \eta(\boldsymbol{a}^t_m), \tag{16}$$

where $\mathcal{P}^a_m$ is the action space of all APs scheduled on channel m, i.e., $\mathcal{P}^a_m = \prod_{k \in \mathcal{K}_m} \mathcal{P}^a_{k,m}$.

The expected regret incurred during the time horizon T_P over all M channels is given by:

$$\bar{R}_p = \sum_{m \in \mathcal{M}} \left\{ T_P \sum_{k \in \mathcal{K}_m} \mu_P(k, m, a^{*P}_{k,m}) - \mathbb{E}\left(\sum_{t=1}^{T_P} \sum_k \mu_P(k, m, a^t_{k,m})\, \eta(\boldsymbol{a}^t_m) \right) \right\}. \tag{17}$$

III. Proposed Solution

A. Proposed Algorithm for the Channel Allocation Problem

Since the time horizon T_C is not necessarily known in advance, the proposed solution, presented in Algorithm 1, proceeds in epochs, each epoch consisting of three phases, namely, *exploration, matching and exploitation*. The exploration phase aims at estimating the previously unknown means of each channel, as well as the number of APs competing for system resources. During this phase, each AP uniformly accesses one channel at a time to estimate its mean reward. AP k accessing channel m gets as feedback the achieved reward on m as well as the total number of APs simultaneously accessing channel m. This phase runs for a constant number of timeslots given by T^0_C. Upon termination, all APs have an estimate $\hat{\boldsymbol{\mu}}_M$ of the means of the channels and of the channel gain experienced over each channel. Each AP also calculates an estimate of the number of APs $\hat{K}$, as was done in [18]. These estimated means and number of APs are used in the second phase of the algorithm where APs play a non-cooperative game with the aim of maximizing the achieved sum rewards. The estimated reward means are taken to be the actual utilities achieved in the matching phase. In other words, after choosing a channel m, if the received reward is non-zero, AP k assumes that this reward is equal to:

$$u_k(m) = \hat{\mu}_M(k, m, k_m). \tag{18}$$

The dynamics of this matching phase, adopted from [34], are described in section III-B. The matching phase runs for $c_1 l^{1+\delta}$ frames, where c_1 and δ are constants and l is the epoch number. The third and final phase is an exploitation phase in which APs settle on the channels that resulted in the best performance in the previous matching phase. The exploitation phase runs for $c_2 2^l$ timeslots, c_2 being a constant.

B. Matching Dynamics

Each AP k is associated with a state $[\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{u}}_k, S]$. The baseline action of AP k is $\bar{\boldsymbol{a}}_k \in \{0, 1\}^{M \times 1}$, such that $\sum_{m=1}^{M} \bar{a}_k(m) = N$. The baseline utility of AP k is

Algorithm 1 Channel Allocation Solution

Initialization: Set $\hat{\mu}_M(k,m,k_m)=0,\ \forall k\in\mathcal{K}, m\in\mathcal{M}, k_m\in[\beta]$. Set $b_k^t=0,\ \forall k\in\mathcal{K}$. Let $\epsilon>0$ and $c\geq KN$.

for $l=1,\ldots,L_C$ **do**

1- Exploration Phase:

  for $t=1:T_C^0$ **do**

    Choose one channel $m\in\mathcal{M}$ uniformly.

    Receive the achieved reward $x_k^t(m)$, and the total number of APs, k_m, accessing channel m simultaneously.

    $W_k^t(m,k_m)=W_k^{t-1}(m,k_m)+x_k^t(m)$, $co_k^t(k_m)=co_k^t(k_m)+1$.

    if $k_m>1$ **then**

      $b_k^t=b_k^{t-1}+1$

    end if

  end for

  Estimate means: $\hat{\mu}_M(k,m,k_m)=\frac{W_k^t(m,k_m)}{co_k^t(k_m)},\forall\ k_m\in[\beta]$.

  Estimate the number of APs according to:

$$\hat{K}=\min\left\{\text{round}\left(\frac{\log\left(\frac{T_C^0-b_k^t}{T_C^0}\right)}{\log\left(1-\frac{1}{M}\right)}+1\right),\beta M\right\}.$$

2- Matching Phase: for the next $c_1 l^{1+\delta}$ frames, play according to the dynamics described in section III-B.

  If $S_k=C$, choose the action to play according to Eq. (20). If $S_k=D$, choose the action according to Eq. (21).

  If the achieved reward for some chosen channel $u_k(m)$, found from Eq. (18), is 0, the AP becomes discontent as per Eq. (23).

  If $\boldsymbol{a}_k\neq\bar{\boldsymbol{a}}_k$ or $\boldsymbol{u}_k\neq\bar{\boldsymbol{u}}_k$ or player k is discontent, the state transition happens according to Eq. (24).

  Each AP keeps a counter of the number of times each action $\boldsymbol{a}'_{\boldsymbol{k}}$ was played and resulted in it being content:

$$F_k^l(\boldsymbol{a}'_{\boldsymbol{k}})=\sum_{t=1}^{c_2 l^{1+\delta}}\mathbb{I}\left(\boldsymbol{a}_k^t=\boldsymbol{a}'_{\boldsymbol{k}},S_k^t=C\right),\tag{19}$$

  with $\mathbb{I}$ being the indicator function.

3- Exploitation phase: for $c_2 2^l$ timeslots:

  Play the action $\boldsymbol{a}_{\boldsymbol{k}}^{\boldsymbol{l}*}=\underset{\boldsymbol{a}_k\in\mathcal{A}_k}{\text{argmax}}\, F_k^l(\boldsymbol{a}_k)$.

end for

$\bar{\boldsymbol{u}}_k$, such that $|\bar{\boldsymbol{u}}_k|=N$. Variable $S\in\{C,D\}$ is the mood of AP k and reflects whether k is content or discontent with the current action and utility. At each frame of the matching phase, each AP chooses an action according to the game dynamics and receives a reward that depends on the collective choices of all the APs. Define $u_{k,\max}=\underset{\boldsymbol{a}}{\text{argmax}}\sum_{m=1}^{M}a_k(m)\mu_M(k,m,k_m)$, where $u_{k,\max}$ is the highest reward achievable by AP k, with a number of estimated APs given by $\hat{K}$.

At each frame t during the matching phase, AP k adheres by the following dynamics to decide on the action to choose:

- A content AP plays its baseline action with high probability:

$$p_k^{\boldsymbol{a}_k}=\begin{cases}\frac{\epsilon^c}{|\mathcal{A}_k|-1}, & \text{if } \boldsymbol{a}_k\neq\bar{\boldsymbol{a}}_k,\\ 1-\epsilon^c, & \text{if } \boldsymbol{a}_k=\bar{\boldsymbol{a}}_k.,\end{cases}\tag{20}$$

  where $\epsilon>0$ is a small perturbation and c is a constant satisfying $c\geq KN$.

- A discontent AP chooses its action uniformly at random:

$$p_k^{\boldsymbol{a}_k}=\frac{1}{|\mathcal{A}_k|},\quad\forall\ \boldsymbol{a}_k\in\mathcal{A}_k.\tag{21}$$

In Eq. (20) and (21), $p_k^{\boldsymbol{a}_k}$ is the probability with which AP k chooses action $\boldsymbol{a}_k$.

After deciding on the action and observing the reward $u_k(m)$ for chosen channels, the state transition of each AP k occurs according to:

- If $\boldsymbol{a}_k=\bar{\boldsymbol{a}}_k$ and $\boldsymbol{u}_k=\bar{\boldsymbol{u}}_k$, a content AP remains content:

$$[\bar{\boldsymbol{a}}_k,\bar{\boldsymbol{u}}_k,C]\rightarrow[\bar{\boldsymbol{a}}_k,\bar{\boldsymbol{u}}_k,C].\tag{22}$$

- If $u_k(m)=0$ for some $m=1,\ldots,N$, AP k becomes discontent with probability one.

$$[\bar{\boldsymbol{a}}_k,\bar{\boldsymbol{u}}_k,C/D]\rightarrow[\boldsymbol{a}_k,\boldsymbol{u}_k,D].\tag{23}$$

- If $\boldsymbol{a}_k\neq\bar{\boldsymbol{a}}_k$ or $\boldsymbol{u}_k\neq\bar{\boldsymbol{u}}_k$ or player k is discontent, the state transitions occur according to:

$$[\bar{\boldsymbol{a}}_k,\bar{\boldsymbol{u}}_k,C/D]\rightarrow\begin{cases}[\boldsymbol{a}_k,\boldsymbol{u}_k,C] & \text{w.p. } \epsilon^{u_{k,\max}-\sum_{n=1}^{N}u_{k,n}},\\ [\boldsymbol{a}_k,\boldsymbol{u}_k,D] & \text{w.p. } 1-\epsilon^{u_{k,\max}-\sum_{n=1}^{N}u_{k,n}}.\end{cases}\tag{24}$$

C. Proposed Solution for the Distributed Power Allocation

A simplified version of Algorithm 1 can be used to solve the power allocation problem over each channel m. The solution is divided into three phases:

1) Exploration phase: This phase runs for T_P^0 timeslots and aims at estimating the reward of each power value. During this phase, each AP chooses each of its possible power levels, i.e., power levels in $\mathcal{P}_k^a$, uniformly at random. Upon termination, APs have estimates of the reward associated to each power value, denoted by $\hat{\boldsymbol{\mu}}_P$.
2) Matching phase: In this phase, APs play a non-cooperative game according to the dynamics presented in Section III-B, after replacing $\mathcal{A}_k$ in Eq. (20) and (21) by $\mathcal{P}_{k,m}^a$. Each AP keeps a counter of the number of times each action was played and resulted in content behavior.
3) Exploitation phase: During this phase, each AP k exploits the action, i.e., the power level, that resulted in the most content behavior during the matching phase.

IV. Regret Analysis

The time horizon of the channel allocation phase can be lower bounded by [31]:

$$T_C\geq\sum_{l=1}^{L_C-1}(T_C^0+c_1 l^{1+\delta}+c_2 2^l)\geq c_2(2^{L_C}-2),\tag{25}$$

where L_C is the total number of epochs occurring within T_C and upper bounded by:

$$L_C\leq\log\left(T_C/c_2+2\right).\tag{26}$$

Similarly, the number of epochs, L_P, occurring within the time horizon T_P dedicated to the power allocation stage is upper bounded by $L_P\leq\log(T_P/c_2+2)$.

A. Regret in the Exploration Phase

In the exploration phase of the channel allocation, each AP samples channels uniformly to get estimates of their means. Even though the purpose of this work is to assign to each AP N channels at each timeslot, the number of channels sampled by each AP at a timeslot is set to one in the exploration phase. The expected regret incurred by all APs in the exploration phase of the channel allocation, R_C^1, can be upper bounded by:

$$R_C^1 \leq \sum_{l=1}^{L_C} KNT_C^0 \leq KNT_C^0 \log\left(T_C/c_2 + 2\right). \tag{27}$$

Similarly, the expected regret incurred by all APs in the exploration phase of the power allocation, R_P^1, can be upper bounded by:

$$R_P^1 \leq \sum_{m=1}^{M} \sum_{l=1}^{L_P} K_m T_P^0 \leq KT_P^0 \log(T_P/c_2 + 2). \tag{28}$$

B. Regret in the Matching Phase

The expected regret in the matching phase of the channel allocation, R_C^2, can be upper bounded by:

$$R_C^2 \leq \sum_{l=1}^{L_C} KNc_1 l^{1+\delta} \leq KNc_1 \log^{2+\delta}\left(T_C/c_2 + 2\right). \tag{29}$$

Similarly, the expected regret in the matching phase of the power allocation, R_P^2, can be upper bounded by:

$$R_P^2 \leq \sum_{m=1}^{M} \sum_{l=1}^{L_C} K_m c_1 l^{1+\delta} \leq Kc_1 \log^{2+\delta}\left(T_P/c_2 + 2\right). \tag{30}$$

C. Regret in the Exploitation Phase

In the exploitation phase of epoch l of the channel allocation, each AP k plays the action that it played the most and resulted in content behavior in the matching phase of epoch l. The exploitation phase fails in two cases:

1) If the exploration phase of epoch l fails: This happens with a probability $\leq 4(M\beta)^2 e^{-l}$ as shown in Lemma 2.
2) If the most played action of the matching epoch differs from the optimal action: This happens with a probability $\leq A_1 e^{-l^{1+\delta}}$ as shown in Lemma 5.

The expected regret incurred by all APs in the exploitation phase can be upper bounded by:

$$R_C^3 \leq \sum_{l=1}^{L_C} KNc_2 2^l \left(4(M\beta)^2 e^{-l} + A_1 e^{-l^{1+\delta}}\right) \leq A_3, \tag{31}$$

where A_1 and A_3 are constants.

Similarly, the regret incurred by the APs in the exploitation phase of the power allocation is $R_P^3 \leq A_3$.

D. Regret of the Proposed Technique

Theorem 1. The expected regret of the proposed allocation solution can be upper bounded as:

$$R \leq R_C^1 + R_C^2 + R_C^3 + R_P^1 + R_P^2 + R_P^3 = \mathcal{O}\left(\log^{2+\delta}(T)\right). \tag{32}$$

V. Exploration Phase

The exploration phase is performed so APs learn estimates of the channel mean reward in the channel allocation phase, and of the power level mean reward in the power allocation phase. Moreover, by keeping track of the number of times each channel was accessed with one or more other APs in the channel allocation phase, the APs can estimate the total number of APs in the system. In this section, we find the minimum length of the exploration phase ensuring an accurate estimation of both the reward means and the number of APs.

A. Estimation of the Reward Means

Since the estimation may not always be perfect, the result of the assignment with the estimated means ($\hat{\boldsymbol{\mu}}_M$ and $\hat{\boldsymbol{\mu}}_P$) might differ from the result of the assignment calculated with the true means ($\boldsymbol{\mu}_M$ and $\boldsymbol{\mu}_P$). However, if the estimation inaccuracy is kept small as in [17] and [31], the result of the assignment would not be affected.

Lemma 1. Let J_M^1 and J_M^2 be the sum reward achieved by the best channel assignment and the second best channel assignment and let $\Delta_M = \frac{J_M^1 - J_M^2}{2KN}$. Moreover, let J_P^1 and J_P^2 be the sum reward achieved by the best power allocation on each channel m and the second best power assignment and let $\Delta_P = \frac{J_P^1 - J_P^2}{2K_m}$. If the difference between the estimated and the correct reward means satisfies:

$$\begin{aligned} |\mu_M(k,m,k_m) - \hat{\mu}_M(k,m,k_m)| < \Delta_M, \\ \forall k \in \mathcal{K}, m \in \mathcal{M}, k_m \in [\beta], \end{aligned} \tag{33}$$

$$\begin{aligned} |\mu_P(k,m,v_l) - \hat{\mu}_P(k,m,v_l)| < \Delta_P, \\ \forall k \in \mathcal{K}_m, m \in \mathcal{M}, v_l \in \mathcal{VL}, \end{aligned} \tag{34}$$

then, the best assignment result does not change due to the estimation inaccuracy.

Proof. See Appendix A. ■

Next, we upper bound the probability of error, i.e., the probability of having channel reward estimates (resp. power level reward estimates) that do not satisfy the condition in (33) (resp. condition (34)) in the exploration epoch l. We also provide a lower bound of the length of the exploration epoch $T_{\hat{\boldsymbol{\mu}}_M}$ in the channel allocation phase, and T_P^0 in the power allocation phase.

Lemma 2. If $T_{\hat{\boldsymbol{\mu}}_M} = \left\lceil \frac{2Me\left(\frac{K-1}{M-1}\right)}{\Delta_M^2 (M-1)^{1-\beta}} \right\rceil$, all players have an estimate of the channel means satisfying the condition in (33), with probability $\geq 1 - \gamma_{e,l}^M$, where $\gamma_{e,l}^M$ is the probability of error in the l^{th} exploration phase of the uncoordinated channel access. Moreover, $\gamma_{e,l}^M \leq 4(M\beta)^2 e^{-l}$.

For the power allocation exploration phase, if $T_P^0 = \left\lceil \frac{2Le\left(\frac{\beta-1}{L-1}\right)}{\Delta_p^2} \right\rceil$, all players have an estimate of the power level means satisfying the condition in (34), with probability $\geq 1 - \gamma_{e,l}^P$, where $\gamma_{e,l}^P$, is the probability of error in the l^{th} exploration phase of the power allocation, upper bounded by $4\beta L e^{-l}$.

Proof. See Appendix B. ■

We now turn our attention to finding the minimum length of the exploration phase in the channel allocation stage ensuring an accurate estimate of the number of APs $\hat{K}$.

B. Estimating the number of APs

For AP k, b_k^t found in step 7 of Algorithm 1 denotes the number of timeslots player k was not the sole occupier of some channel m until t.

Lemma 3. If the length of the exploration epoch in the channel allocation step satisfies:

$$T_{\hat{K}} = \left\lceil 2.08 \log\left(\frac{2}{\eta}\right) M^2 e^{2\left(\frac{M\beta-1}{M-1}\right)} \right\rceil, \tag{35}$$

then all APs have an estimate of the number of APs $\hat{K}$ satisfying $\hat{K} = K$ with probability higher than $1-\eta$, where η is the probability of error in the estimation of the number of APs.

Proof. See Appendix C. ■

C. Length of the Channel Allocation Exploration Phase

To ensure an accurate estimate of the channel reward means and of the number of APs, the minimum length of the exploration phase in the channel allocation solution, T_C^0, must satisfy the conditions in Lemma 2 and Lemma 3. Hence, the following must hold:

$$T_C^0 = \max\left\{ \left\lceil \frac{2Me^{\left(\frac{K-1}{M-1}\right)}}{\Delta_M^2 \left(M-1\right)^{1-\beta}} \right\rceil, \left\lceil 2.08 \log\left(\frac{2}{\eta}\right) M^2 e^{2\left(\frac{M\beta-1}{M-1}\right)} \right\rceil \right\}. \tag{36}$$

VI. Matching Phase

The matching phase of the channel allocation solution aims at reaching a final assignment in which every AP accesses N channels, such that the achieved sum reward is maximized.

The dynamics presented in section III-B and adopted in the matching phase induce a Markov chain over the state space $\mathcal{Z} = \prod_{k=1}^{K} \{\mathcal{A}_K \times [0,1]^{N\times 1} \times \{C, D\}\}$. Let P^ϵ denote the transision matrix of the regular perturbed Markov chain $\mathcal{Z}$. The work in [34] guarantees that, when playing according to these dynamics, the optimal state, i.e., the one maximizing the sum rewards, is played most often. The proof relies on the theory of resistance trees for regular perturbed Markov chains [35]. The dynamics used in this paper differ from those in [34] in two aspects:

1) If AP k receives a reward equal to 0 on some channel m, AP k is discontent with probability one. In [34], the game is assumed to be interdependent which means that it is not possible to partition APs into two groups that do not interact with each other. However, this property does not hold in the considered setting as shown in [31]. Therefore, as in [31], to characterize the stable states of the unperturbed chain when $\epsilon = 0$, a player with 0 reward on some channels is discontent with probability one.
2) For the transition probabilities between content and discontent in Eq. (24), instead of using $\epsilon^{N-\sum_{n=1}^{N} u_{k,n}}$, we use $\epsilon^{u_{k,\max}-\sum_{n=1}^{N} u_{k,n}}$, since the maximum utility achievable by each AP k is $u_{k,\max}$.

Next, the recurrence states of $\mathcal{Z}$ are characterized.

Lemma 4. Let D^0 denote the set of states where all APs are discontent. Moreover, let C^0 denote all singleton states where all APs are content and their baseline actions and utilities are aligned. As proved in [34], the only recurrence states of $\mathcal{Z}$ are D^0 and all singletons in C^0.

The resistance of moving from one recurrence state to the other being similar to [34], the stochastic potential of any state $z \in C^0$ is of the form:

$$\zeta(z) = c[|C^0| - 1] + \sum_{k=1}^{K} u_{k,\max} - \sum_{m=1}^{M} a_k(m)\hat{\mu}(k, m, k_m). \tag{37}$$

From Theorem 1 of [34], the stable state is the one minimizing the stochastic potential, hence the one maximizing the achieved sum reward. This stable state is guaranteed to be played the majority of times for a small enough perturbation ϵ [31], [34]. In the exploitation phase, as the state that was most played and that resulted most in the players being content is played, the stable state is hence expected to be played with high probability. Next, the probability of error in the matching epoch l is found.

Let π denote the stationary distribution of the Markov chain $\mathcal{Z}$ and let $\boldsymbol{z}^* = [\bar{\boldsymbol{a}}^*, \bar{\boldsymbol{u}}^*, C^K]$ denote the optimal state. According to [31], $\pi(\boldsymbol{z}^*) > 1/2$ for a small enough perturbation ϵ. The following lemma finds the probability of error in the matching phase of the l^{th} epoch, $\delta_{m,l}$.

Lemma 5. Let $\boldsymbol{a}^{(l)}$ denote the action that was most played in some epoch l. As proved in [17], the probability of error in the matching phase in epoch l, $\delta_{m,l}$, is upper bounded by:

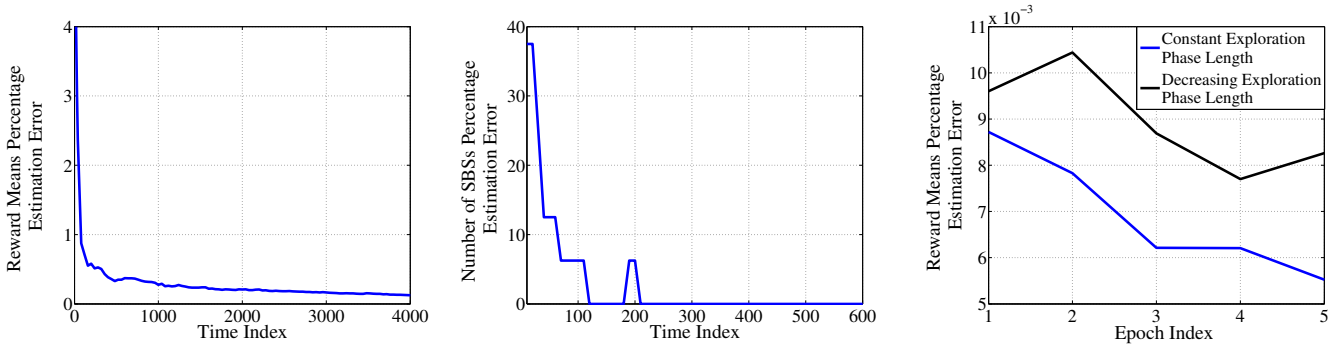
$$\delta_{m,l} = \Pr(\boldsymbol{a}^* \neq \boldsymbol{a}^{(l)}) \leq A_0 \left\|\phi\right\|_\pi \exp\left(\frac{-\theta^2 \pi(\boldsymbol{z}^*) c_2 l^{1+\delta}}{72 T_m(1/8)}\right), \tag{38}$$

where A_0 is a constant, ϕ_π is the probability distribution of the initial state played in epoch l and $T_m(1/8)$ is the mixing time of the Markov chain $\mathcal{Z}$ with an accuracy of 1/8 [36].

The analysis of the matching phase of the power allocation solution is similar to the one given above and is omitted for space constraints.

VII. Simulation Results

Extensive simulations of the proposed algorithm were conducted to validate its performance. The following simulation parameters were chosen: $K = 4, M = 4, N = \beta = L = 2, B_c = 2.5$ MHz, $c_1 = 3000, c_2 = 5000, \epsilon = 5 \times 10^{-5}, \gamma = 0$. The available SINR values are $\mathbf{\Gamma} = \{24, 4.77\}$ (dB) leading to achieved rates of 20 and 5 Mbps respectively. For the channel allocation stage, the parameter c used in the matching phase (Cf. Section III-B) is set as: $c = KN$, whereas for the power allocation stage $c = K_m$ for each channel $m \in \mathcal{M}$. Two of the APs are assumed to have a power budget of 1W per



(a) (b) (c)

Fig. 2: Estimation error as time progresses in the channel allocation stage for (a) the estimation of the rewards, (b) the estimation of the number of APs. (c) Comparison of the estimation error as a function of the epoch index in the channel allocation stage for the estimation of the rewards.

channel, while the remaining two have a power budget of 2W per channel. Additional simulation parameters are given in Table I [37].

TABLE I: Simulation parameters.

Cell Radius R_d	150 m
Overall Transmission Bandwidth	10 MHz
Number of channels	4
Number of APs	4
Power Budget per AP per channel $P^m_{(.)}$	$\{1, 1, 2, 2\}$ (W)
Available SINR Requirements	$\mathbf{\Gamma} = \{24, 4.77\}$(dB)
Distance Dependent Path Loss	$128.1 + 37.6 \log_{10}(d)$(dB), d in Km
Receiver Noise Density	4.10^{-18} mW/Hz

A. Estimation Accuracy of the Exploration Phase

First, we evaluate the estimation accuracy of the exploration phase in the channel allocation stage. As shown in Fig. 2a and Fig. 2b, the estimation of both the reward means and the total number of APs converges rather quickly to the correct values. Having observed that the estimation of the exploration phase converges quickly, a version of the proposed algorithm where the exploration phase length is divided by the epoch index was tested. The estimation error of this version with a decreasing exploration phase length was compared against the version with a constant exploration phase length. Fig. 2c plots the channel rewards estimation error for both versions. Although the constant length version outperforms the version with a decreasing exploration phase length, the estimation error achieved by both methods is lower than $1.1 \times 10^{-2}\%$, hence negligible. When it comes to the number of APs estimation, both versions accurately estimate $\hat{K}$, without error, when convergence is reached.

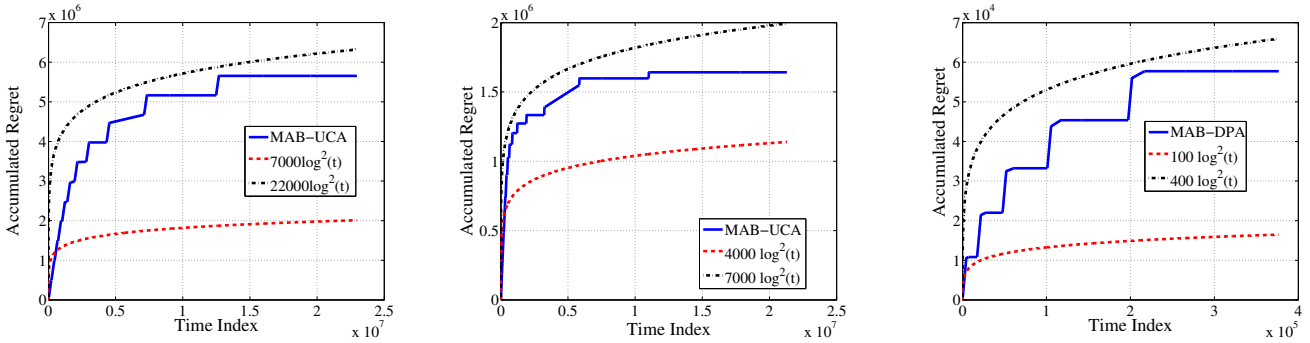
For the power allocation stage, the power level rewards estimation also converges quickly to a negligible error value.

B. Performance Analysis

Fig. 3 shows the average accumulated regret as a function of time in the channel allocation stage for both the constant and the decreasing length exploration phase versions. The results show that the average accumulated regret for both versions increases with time as $\mathcal{O}(\log(t)^2)$. More specifically, the regret incurred for the constant length exploration phase version is bounded between $7000 \log(t)^2$ and $22000 \log(t)^2$, as shown in Fig. 3a. The regret incurred for the decreasing length exploration phase version is bounded between $4000 \log(t)^2$ and $7000 \log(t)^2$. In fact, most of the regret is accumulated during the exploration phase where APs choose a channel uniformly at random. Hence, decreasing the length of the exploration phase lowers the value of the accumulated regret as shown in Fig. 3b, without jeopardizing the estimation accuracy as was shown in Section VII-A.

The regret incurred on all channels during the power allocation stage is bounded between $100 \log(t)^2$ and $400 \log(t)^2$, as shown in Fig. 3c. The lower regret observed during the power allocation stage, when compared to the channel allocation stage, results from the smaller number of APs competing for a smaller number of arms. In fact, on each channel $m \in \mathcal{M}$ during the power allocation stage, the number of competing APs is $K_m \leq \beta = 2$, while the number of arms or power levels is $L = 2$. In contrast, during the channel allocation stage, the number of players is $K = 4$ with $\binom{M}{N} = 6$ available arms.

Remark. To provide insight on the accumulated regret as a function of time in seconds, and the time duration needed to reach convergence, assume that a subcarrier spacing of 240 KHz [38] is considered, resulting in a timeslot duration equal to 62.5 μs. In Fig. 3a and 3b, the performance of the proposed solution is evaluated for a large number of timeslots to assess its performance in the long run. Therefore, in Fig. 4, the accumulated regret in the uncoordinated channel access part of the solution is plotted for a shorter duration. Fig. 4 shows that convergence to the optimal allocation is first reached at the fourth epoch, which takes place from 0.45×10^6 to 0.6×10^6 timeslots approximately. In terms of time duration in seconds, convergence is reached in $0.45 \times 10^6 \times 62.5 \times 10^{-6} = 28.125$ seconds. From the subsequent epochs shown by Fig. 4, the system converges to the optimal allocation in each epoch, resulting in zero regret in the matching and the exploitation



(a) (b) (c)

Fig. 3: Accumulated regret as time progresses (a) for the channel allocation phase with a constant exploration phase length, (b) for the channel allocation phase with a decreasing exploration phase length, (c) for the power allocation stage.

phases. Note that for the uncoordinated power control part, convergence is reached from the first epoch, i.e., at around 0.1×10^5 timeslots, or 0.625 seconds with a timeslot duration of 62.5 μs.

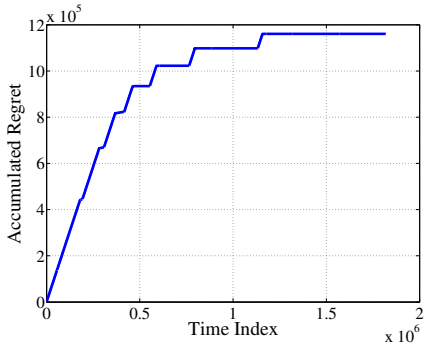


Fig. 4: Accumulated regret as a function of time for $K = M = 4$.

In Fig. 5, we compare the performance of the proposed method with a technique based on the UCB algorithm proposed in [24] and similar to the one proposed in [23], denoted by Two-Dimensional UCB. In the Two-Dimensional UCB method, channel and power allocation are conducted at the same time, using the UCB algorithm, by considering all possible combinations of the channels and the power levels. For the considered setting, the number of arms in the Two-Dimensional UCB method is hence $\binom{M}{N} \times L^N = 24$ arms. In Fig. 5a, the achieved rate is plotted as a function of time. Both methods converge relatively quickly to the highest achievable rate, with small variations for the Two-Dimensional UCB technique. The sharp falls in the achieved rate of the proposed method are due to the exploration phase during each epoch of the power allocation stage where APs choose the power levels uniformly at random, causing collisions and leading to zero rates.

The total transmit power used by the APs as a function of time is shown in Fig. 5b. While both methods converge to the same highest achievable rate, the power used by our proposed method is significantly lower than the one needed by the Two-Dimensional UCB method. This means that the UCB-based method does not lead APs to learn the optimal allocation and converges to a sub-optimal resource partitioning among the APs. In other words, our proposed method achieves a better allocation for the channel and power when compared to the UCB-based method. Moreover, our proposed method has performance guarantees in terms of regret and optimality, while the Two-Dimensional UCB method [24] does not.

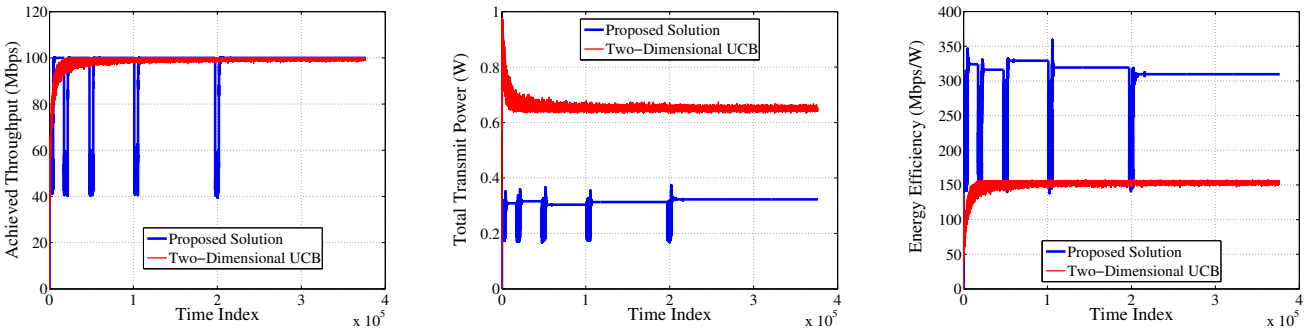
To check the combined effect of rate and power on the performance of the compared methods, the achieved energy efficiency (EE), which is the ratio of the achieved rate to the used power, is plotted in Fig. 5c. Once again, the sharp falls in the performance of our proposed method are due to the exploration phase in each epoch of the power allocation stage. Fig. 5c shows that our proposed method greatly outperforms the UCB-based method, by achieving more than a twofold increase in the EE. This is due to our method converging to the optimal allocation when the UCB-based technique converges to a sub-optimal allocation requiring more transmit power as shown by Fig. 5b.

VIII. Conclusion

In this paper, the uncoordinated channel and power allocation problems in a SON were studied. The considered framework allows each AP to choose N channels at each timeslot, and allows each channel to simultaneously accommodate multiple APs in a NOMA manner. The considered problem was modeled using the multi-player MAB framework, with varying user rewards, multiple plays, and non-zero reward on collision. A game-theoretic approach was used to develop an algorithm with a sub-linear regret of $\mathcal{O}(\log^2 T)$. Simulation results validated the sub-linear regret of the proposed method and showed its superior performance, when compared with one of the most used algorithms in the MAB literature.

Acknowledgment

The first author would like to thank Akshayaa Magesh for useful discussions regarding multi-player multi-armed bandits.



(a) (b) (c)

Fig. 5: Performance comparison as a function of time of (a) the achieved rate, (b) the total transmit power, (c) the energy efficiency.

APPENDIX A
PROOF OF LEMMA 1

In the channel allocation phase, denote by $\boldsymbol{a}^{(1)}$ the optimal assignment, and by J_M^1 the sum rewards achieved when $\boldsymbol{a}^{(1)}$ is played, which is then given by:

$$J_M^1 = \sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)\mu_M(k,m,k_m^*). \tag{39}$$

Furthermore, denote the second best assignment and the sum reward achieved under it by $\boldsymbol{a}^{(2)}$ and J_M^2 respectively. Let the estimated mean of AP k over channel m with k_m APs on channel m be written as:

$$\hat{\mu}_M(k,m,k_m) = \mu_M(k,m,k_m) + z(k,m,k_m), \tag{40}$$

where $z(k,m,k_m)$ is the estimation inaccuracy during the channel allocation phase satisfying $|z(k,m,k_m)| \leq \Delta_M$. The sum reward achieved when $\boldsymbol{a}^{(1)}$ is played with the estimated channel means satisfies:

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)\hat{\mu}_M(k,m,k_m) = \\ &\sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)(\mu_M(k,m,k_m) + z(k,m,k_m)) > \\ &\sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)\mu_M(k,m,k_m) - KN\Delta_M.\end{aligned} \tag{41}$$

Any other assignment $\boldsymbol{a} \neq \boldsymbol{a}^{(1)} \neq \boldsymbol{a}^{(2)}$ must perform at most as well as $\boldsymbol{a}^{(2)}$:

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)\hat{\mu}_M(k,m,k_m) = \\ &\sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)(\mu_M(k,m,k_m) + z(k,m,k_m)) < \\ &\sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(2)}(m)\mu_M(k,m,k_m) + KN\Delta_M.\end{aligned} \tag{42}$$

To avoid changing the optimal assignment because of the estimation inaccuracy, the following must hold $\forall \boldsymbol{a} \neq \boldsymbol{a}^{(1)}$:

$$\sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)\hat{\mu}_M(k,m,k_m) > \sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)\hat{\mu}_M(k,m,k_m). \tag{43}$$

To ensure Eq. (43), we need to have: $J_M^1 - KN\Delta_M > J_M^2 + KN\Delta_M$, which holds if:

$$\Delta_M < \frac{J_M^1 - J_M^2}{2KN}. \tag{44}$$

In the power allocation phase, following a similar approach over each channel m, we get:

$$\Delta_P < \frac{J_P^1 - J_P^2}{2K_m}. \tag{45}$$

APPENDIX B
PROOF OF LEMMA 2

A. Lower Bound of the Length of the Exploration Phase in the Channel Allocation Step

To find a lower bound of the length of the exploration phase in the channel allocation step, we first find the required number of observations of each channel by each AP to guarantee condition (33) [18], [39]. To do so, the probability of each AP not having a correct estimation of the channel means should be bounded. Let $\gamma = \gamma_{e,l}^M/2$. Define the following events:

- A: all players have an estimate satisfying condition (33),
- B: all players have $\geq Q$ observations of each channel m for every s in $[\beta]$,
- A_k: player k has an estimate satisfying condition (33),
- B_k: player k has $\geq Q$ observations of each channel m for every s in $[\beta]$.

The following must hold:

$$\Pr(\bar{A}_k \mid B_k) \leq \frac{\gamma}{K}. \tag{46}$$

In fact,

$$\begin{aligned}&\Pr(\bar{A}_k|B_k) \leq \\ &\Pr\left(\exists\ m,s, \text{s.t. } |\mu_M(k,m,s)-\hat{\mu}_M(k,m,s)|>\Delta_M \mid B_k\right) \overset{(a)}{\leq} \\ &\sum_{m=1}^{M}\sum_{s=1}^{\beta}\Pr\left(|\mu_M(k,m,s)-\hat{\mu}_M(k,m,s)|>\Delta_M \mid B_k\right) = \\ &\sum_{m=1}^{M}\sum_{s=1}^{\beta}\sum_{q=Q}^{\infty}\Pr\left(|\mu_M(k,m,s)-\hat{\mu}_M(k,m,s)|>\Delta_M \mid \right. \\ &\left. k \text{ has } q \text{ observations of } (m,s)\right) \times p_2 \overset{(b)}{\leq} \\ &\sum_{m=1}^{M}\sum_{s=1}^{\beta}\sum_{q=Q}^{\infty}2p_2e^{(-2q\Delta_M^2)} \leq \sum_{m=1}^{M}\sum_{s=1}^{\beta}2e^{(-2Q\Delta_M^2)} = \\ &2M\beta e^{(-2Q\Delta_M^2)},\end{aligned} \tag{47}$$

where (m,s) refers to channel m with s players on it, (a) results from applying the union bound and (b) from using Hoeffding's inequality [40], and $p_2 = \Pr\left(q \text{ observations of } (m,s) \mid q \geq Q\right)$.

To ensure $\Pr(\bar{A}_k|B_k)$ is lower than $\frac{\gamma}{K}$, Q must satisfy:

$$Q \geq \frac{1}{2\Delta_M^2}\log(\frac{2KM\beta}{\gamma}). \tag{48}$$

Then,

$$\Pr\left(A|B\right) = 1-\Pr(\bar{A}|B) \geq 1-\sum_{k=1}^{K}\Pr\left(\bar{A}_k|B_k\right) = 1-\gamma, \tag{49}$$

leading to all APs having an estimate of every channel satisfying condition (33) with probability higher than $1-\gamma$.

Next, we need to find a time horizon T_h for the exploration phase of the channel allocation step large enough such that all players have $\geq Q$ observations of each arm with probability higher than $1-\gamma$. Note that the length of each exploration phase $T_{\hat{\boldsymbol{\mu}}}$ does not necessarily satisfy $T_{\hat{\boldsymbol{\mu}}} \geq T_h$. In other words, all players can get $\geq Q$ observations of each arm with probability higher than $1-\gamma$ after multiple exploration phases.

Let $A_{k,m,s}(t)=1$ if player k observed channel m with s APs on it at timeslot t, and 0 otherwise. For $0<\tau<1$, we have:

$$\begin{aligned}&\Pr\left(\ k \text{ has } \leq (1-\tau)T_h\mathbb{E}[A_{k,m,s}] \text{ observations}\right) = \\ &\Pr\left(\sum_{t=1}^{T_h}A_{k,m,s}(t) \leq (1-\tau)T_h\mathbb{E}[A_{k,m,s}]\right) = \\ &\Pr\left(e^{\left(-d\sum_{t=1}^{T_h}A_{k,m,s}(t)\right)} \geq e^{(-d(1-\tau)T_h\mathbb{E}[A_{k,m,s}])}\right) \overset{(a)}{\leq} \\ &\frac{\mathbb{E}\left[e^{\left(-d\sum_{t=1}^{T_h}A_{k,m,s}(t)\right)}\right]}{e^{(-d(1-\tau)T_h\mathbb{E}[A_{k,m,s}])}},\end{aligned} \tag{50}$$

where $d>0$ and (a) results from applying the Chernoff bound. By noting that all players are randomly and uniformly sampling every channel during the exploration phase, for any $k \in \mathcal{K}, m \in \mathcal{S}, s \in [\beta]$, $A_{k,m,s}$ are i.i.d. across time. Hence:

$$\mathbb{E}\left[e^{\left(-d\sum_{t=1}^{T_h}A_{k,m,s}(t)\right)}\right] = \prod_{t=1}^{T_h}\mathbb{E}\left[e^{(-dA_{k,m,s}(t))}\right]. \tag{51}$$

Moreover, $A_{k,m,s}(t)$ is a Bernoulli random variable that takes the value 1 with probability p_A. Therefore, we have:

$$\mathbb{E}\left[e^{(-dA_{k,m,s}(t))}\right] = 1+p_A(e^{-d}-1) \overset{(a)}{\leq} e^{\left(p_A\left(e^{-d}-1\right)\right)}, \tag{52}$$

where (a) follows since $1+y \leq e^y$. Eq. (51) can hence be expressed as:

$$\begin{aligned}\mathbb{E}\left[e^{\left(-d\sum_{t=1}^{T_h}A_{k,m,s}(t)\right)}\right] \leq e^{\sum_{t=1}^{T_h}\left(p_A\left(e^{-d}-1\right)\right)} = \\ e^{\left(T_h\mathbb{E}[A_{k,m,s}]\left(e^{-d}-1\right)\right)}.\end{aligned} \tag{53}$$

By inserting Eq. (53) into Eq. (50), we get:

$$\begin{aligned}&\Pr\left(\text{ player } k \text{ has } \leq (1-\tau)T_h\mathbb{E}[A_{k,m,s}]\ \right) \leq \\ &e^{\left(T_h\mathbb{E}[A_{k,m,s}]\left(e^{-d}-1\right)\right)+(d(1-\tau)T_h\mathbb{E}[A_{k,m,s}])}.\end{aligned} \tag{54}$$

To make the bound as tight as possible, d is chosen such that the right hand side of Eq. (54) is minimized, leading to $d=-\log(1-\tau)$. By substituting d by its value in Eq. (54), we get:

$$\begin{aligned}&\Pr\left(\text{ player } k \text{ has } \leq (1-\tau)T_h\mathbb{E}[A_{k,m,s}]\ \right) \leq \\ &e^{(-T_h\mathbb{E}[A_{k,m,s}](\tau-(1-\tau)\log(1-\tau)))} = \\ &\left(\frac{e^{-\tau}}{(1-\tau)^{(1-\tau)}}\right)^{(T_h\mathbb{E}[A_{k,m,s}])} \overset{(a)}{\leq} e^{-\frac{\tau^2}{2}T_h\mathbb{E}[A_{k,m,s}]},\end{aligned} \tag{55}$$

where (a) results from having $(1-\tau)\log(1-\tau) > -\tau+\frac{\tau^2}{2}$, obtained by using a Taylor expansion.

Taking $\tau=1/2$ and using a union bound on (55), we get:

$$\begin{aligned}&\Pr\left(\exists\ k,m,s \text{ s.t. } k \text{ has } \leq \frac{T_h}{2}\mathbb{E}[A_{k,m,s}(t)] \text{ observations}\right) \leq \\ &KM\beta e^{\left(\frac{-\frac{1}{4}T_h\mathbb{E}[A_{k,m,s}]}{2}\right)},\end{aligned} \tag{56}$$

which is upper bounded by γ if T_h satisfies:

$$T_h \geq \frac{8}{\mathbb{E}[A_{k,m,s}]}\log\left(\frac{KM\beta}{\gamma}\right). \tag{57}$$

Moreover, the number of observations of each arm during T_h timeslots, $\sum_{t=1}^{T_h}A_{k,m,s}(t)$, must be at least equal to Q. Hence we need:

$$\sum_{t=1}^{T_h}A_{k,m,s}(t) > \frac{T_h}{2}\mathbb{E}[A_{k,m,s}] \geq Q > \frac{1}{2\Delta_M^2}\log\left(\frac{2KM\beta}{\gamma}\right), \tag{58}$$

which holds if:

$$\begin{aligned}T_h \geq \left\lceil \max\left\{\frac{8}{\mathbb{E}[A_{k,m,s}]}\log\left(\frac{KM\beta}{\gamma}\right),\right.\right. \\ \left.\left.\frac{1}{\Delta_M^2\mathbb{E}[A_{k,m,s}]}\log\left(\frac{2KM\beta}{\gamma}\right)\right\}\right\rceil.\end{aligned} \tag{59}$$

Note that:

$$\mathbb{E}[A_{k,m,s}] = \binom{K-1}{s-1}\left(\frac{1}{M}\right)^{s}\left(1-\frac{1}{M}\right)^{K-s} \overset{(a)}{\geq} \left(\frac{1}{M}\right)^{s}\left(1-\frac{1}{M}\right)^{K-s} \geq \left(\frac{1}{M}\right)\left(\frac{1}{M}\right)^{s-1}\left(1-\frac{1}{M}\right)^{K-1}\left(1-\frac{1}{M}\right)^{1-s} \overset{(b)}{\geq} \frac{1}{Me^{\left(\frac{K-1}{M-1}\right)}}(M-1)^{1-s} \overset{(c)}{\geq} \frac{(M-1)^{1-\beta}}{Me^{\left(\frac{K-1}{M-1}\right)}}, \tag{60}$$

where (a) follows from having $\binom{K-1}{s-1} \geq 1$, (b) from the fact that $(1-\frac{1}{x})^{x-1} \geq \frac{1}{e}$, and (c) from $s \leq \beta$.

Hence, T_h can be re-written as:

$$T_h \geq \left\lceil \max\left\{ \frac{8Me^{\left(\frac{K-1}{M-1}\right)}}{(M-1)^{1-\beta}}\log\left(\frac{KM\beta}{\gamma}\right), \frac{Me^{\left(\frac{K-1}{M-1}\right)}}{\Delta_M^2(M-1)^{1-\beta}}\log\left(\frac{2KM\beta}{\gamma}\right)\right\}\right\rceil. \tag{61}$$

Having T_h, the probability of all APs having an estimate of the channel means satisfying Eq. (33) is lower bounded by:

$$\Pr(A) = 1-\Pr(\bar{A}) = 1-\big(\Pr(\bar{A}|B)\Pr(B)+\Pr(\bar{A}|\bar{B})\Pr(\bar{B})\big) \geq 1-\big(\Pr(\bar{A}|B)\big)+\Pr(\bar{B})\big) \geq 1-(\gamma+\gamma) = 1-\gamma_{e,l}^{M}. \tag{62}$$

Since $\Delta_M = \frac{J_M^1 - J_M^2}{2KN} \leq \frac{KN-0}{2KN} \leq \frac{1}{2}$, Eq. (61) is satisfied if:

$$T_h = \frac{2Me^{\left(\frac{K-1}{M-1}\right)}}{\Delta_M^2(M-1)^{1-\beta}}\log\left(\frac{4KM\beta}{\gamma_{e,l}^{M}}\right). \tag{63}$$

Having found the minimum needed length of the exploration epoch in the channel allocation phase, next, we upper bound the error probability in the l^{th} exploration epoch. To do so, we first note that:

$$T_{\hat{\boldsymbol{\mu}}_{\boldsymbol{M}}} \times l = T_h = \frac{2Me^{\left(\frac{K-1}{M-1}\right)}}{\Delta_M^2(M-1)^{1-\beta}}\log\left(\frac{4KM\beta}{\gamma_{e,l}^{M}}\right). \tag{64}$$

To have $\gamma_{e,l}^{M} \leq 4KM\beta e^{-l} \leq 4(M\beta)^2 e^{-l}$, the length of each exploration epoch must satisfy:

$$T_{\hat{\boldsymbol{\mu}}_{\boldsymbol{M}}} \geq \frac{2Me^{\left(\frac{K-1}{M-1}\right)}}{\Delta_M^2(M-1)^{1-\beta}}. \tag{65}$$

B. Lower Bound of the Length of the Exploration Phase in the Power Allocation Step

By following a similar analysis of the one in Appendix B-A, the minimum length of the length of the exploration phase on each channel m in the power allocation step can be given by:

$$T_P^0 = \left\lceil \frac{2Le^{\left(\frac{\beta-1}{L-1}\right)}}{\Delta_p^2} \right\rceil. \tag{66}$$

If the length of the exploration phase in the power allocation step on each channel m satisfies Eq. (66), then all players have an estimate of the power level means satisfying the condition in (34), with probability $\geq 1-\gamma_{e,l}^{P}$, where $\gamma_{e,l}^{P}$ is upper bounded by $4\beta Le^{-l}$.

APPENDIX C
PROOF OF LEMMA 3

Let p be the true probability of player k not being the sole occupier of some channel m when k accesses the M channels uniformly at random:

$$p = 1-\sum_{m=1}^{M}\frac{1}{M}\left(1-\frac{1}{M}\right)^{K-1} = 1-\left(1-\frac{1}{M}\right)^{K-1}. \tag{67}$$

From Eq. (67), the number of APs K is given by:

$$K = \text{round}\left(\frac{\log(1-p)}{\log(1-\frac{1}{M})}+1\right). \tag{68}$$

The estimated probability of player k not accessing channel m alone at time t is: $\hat{p}_t = b_k^t/t$. For a correct estimation of the number of APs, we need to find a time t sufficiently large to guarantee with high probability that:

$$\hat{K} = \text{round}\left(\frac{\log(1-\hat{p}_t)}{\log(1-\frac{1}{M})}+1\right) = \text{round}\left(\frac{\log(1-p)}{\log(1-\frac{1}{M})}+1\right) = K. \tag{69}$$

To ensure Eq. (69), if $\kappa < 1/2$, the following must hold:

$$\left|\frac{\log(\frac{t-b_k^t}{t})}{\log(1-\frac{1}{M})} - \frac{\log(1-p)}{\log(1-\frac{1}{M})}\right| = \left|\frac{\log\left(\frac{1-\hat{p}_t}{1-p}\right)}{\log\left(1-\frac{1}{M}\right)}\right| \leq \kappa. \tag{70}$$

Let $\hat{p}_t - p = \xi$. After some calculations, Eq. (70) can be expressed as:

$$(1-p)\left(1-\left(1-\frac{1}{M}\right)^{-\kappa}\right) \leq \xi \leq (1-p)\left(1-\left(1-\frac{1}{M}\right)^{\kappa}\right). \tag{71}$$

With high probability, $K = \hat{K}$ when $\kappa < \frac{1}{2}$, if $|\hat{p}_t - p| \leq \xi_1$, where:

$$\xi_1 = \min\left\{\left|(1-p)\left(1-\left(1-\frac{1}{M}\right)^{-\kappa}\right)\right|, \left|(1-p)\left(1-\left(1-\frac{1}{M}\right)^{\kappa}\right)\right|\right\}. \tag{72}$$

Let $T_{\hat{K}}$ be a large enough time horizon for which the estimated probability $\hat{p}_{T_{\hat{K}}}$ is an average of i.i.d. random variables with expectation p. Using Hoeffding's inequality [40], we get:

$$\Pr\left(|\hat{p}_{T_{\hat{K}}} - p| \geq \xi_1\right) \leq 2e^{-2T_{\hat{K}}\xi_1^2}. \tag{73}$$

To bound the probability of an incorrect estimation of $\hat{K}$ by some small value η, $T_{\hat{K}}$ must be lower bounded by:

$$T_{\hat{K}} \geq \frac{\log(2\eta)}{2\xi_1^2}. \tag{74}$$

To get a simpler expression of ξ_1 and hence of $T_{\hat{K}}$, suppose that $\kappa = 0.49$. With the expression of p given by Eq. (67), the

first term in Eq. (72) can be lower bounded as:

$$\begin{aligned}&\left|\left(1-\frac{1}{M}\right)^{K-1}\left(1-\left(1-\frac{1}{M}\right)^{-0.49}\right)\right| = \\ &-\left(1-\frac{1}{M}\right)^{K-1}\left(1-\left(1-\frac{1}{M}\right)^{-0.49}\right)\overset{(a)}{\geq} \\ &\left(1-\frac{1}{M}\right)^{M\beta-1}\left(1-\left(-1+\frac{1}{M}\right)^{-0.49}\right)\overset{(b)}{\geq} \\ &\frac{1}{e^{\left(\frac{M\beta-1}{M-1}\right)}}\left(1-\left(-1+\frac{1}{M}\right)^{-0.49}\right)\overset{(c)}{\geq}\frac{0.49}{Me^{\left(\frac{M\beta-1}{M-1}\right)}},\end{aligned} \tag{75}$$

where (a) results from having $M\beta \geq K$, (b) from $(1-\frac{1}{x})^{x-1} \geq \frac{1}{e}$, and (c) from using a Taylor Expansion. Similarly, the second term in Eq. (72) can be lower bounded as:

$$\left|\left(1-\frac{1}{M}\right)^{K-1}\left(1-\left(1-\frac{1}{M}\right)^{0.49}\right)\right| \geq \frac{0.49}{Me^{\left(\frac{M\beta-1}{M-1}\right)}}. \tag{76}$$

Variable ξ_1 is therefore lower bounded by:

$$\xi_1 \geq \frac{0.49}{Me^{\left(\frac{M\beta-1}{M-1}\right)}}. \tag{77}$$

Hence, $\hat{K} = K$ with probability higher than $1-\eta$ if:

$$T_{\hat{K}} = \left\lceil 2.08\log(2/\eta)M^2e^{2\left(\frac{M\beta-1}{M-1}\right)}\right\rceil. \tag{78}$$

REFERENCES

[1] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.

[2] H. Elsawy, E. Hossain, and D. I. Kim, "HetNets with cognitive small cells: user offloading and distributed channel access techniques," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 28–36, June 2013.

[3] S. Sesia, I. Toufik, and M. Baker, *LTE: The UMTS Long Term Evolution, From Theory to Practice*. New York, USA: Wiley, 2009.

[4] R. Razavi and H. Claussen, "Urban small cell deployments: Impact on the network energy consumption," in *Proc. IEEE Wireless Commun. and Networking Conf. (WCNC)*, Paris, France, Apr. 2012, pp. 47–52.

[5] J. Farah, A. Kilzi, C. Abdel Nour, and C. Douillard, "Power Minimization in Distributed Antenna Systems Using Non-Orthogonal Multiple Access and Mutual Successive Interference Cancellation," *IEEE Trans. on Veh. Technol.*, vol. 67, no. 12, pp. 11 873–11 885, Dec. 2018.

[6] M. Rahman and H. Yanikomeroglu, "Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1414–1425, Apr. 2010.

[7] A. Bin Sediq, R. Schoenen, H. Yanikomeroglu, and G. Senarath, "Optimized distributed inter-cell interference coordination (ICIC) scheme using projected subgradient and network flow optimization," *IEEE Trans. Commun.*, vol. 63, no. 1, pp. 107–124, Jan. 2015.

[8] J. Yun and K. G. Shin, "Distributed coordination of co-channel femtocells via inter-cell signaling with arbitrary delay," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1127–1139, June 2015.

[9] M. Yassin, Y. Dirani, M. Ibrahim, S. Lahoud, D. Mezher, and B. Cousin, "A novel dynamic inter-cell interference coordination technique for LTE networks," in *Proc. IEEE Annual Int. Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Hong Kong, China, Sept. 2015, pp. 1380–1385.

[10] O. Iacoboaiea, B. Sayrac, S. Ben Jemaa, and P. Bianchi, "SON Coordination in Heterogeneous Networks: A Reinforcement Learning Framework," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5835–5847, Sept. 2016.

[11] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge: Cambridge Univ. Press, 2020.

[12] A. Feki and V. Capdevielle, "Autonomous resource allocation for dense LTE networks: A multi armed bandit formulation," in *Proc. IEEE Annual Int. Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Toronto, ON, Canada, Sept. 2011, pp. 66–70.

[13] A. Feki, V. Capdevielle, and E. Sorsy, "Self-organized resource allocation for LTE pico cells: A reinforcement learning approach," in *Proc. IEEE Veh. Techn. Conf. Spring (VTC)*, Yokohama, Japan, May 2012, pp. 1–5.

[14] P. Coucheney, K. Khawam, and J. Cohen, "Multi-armed bandit for distributed inter-cell interference coordination," in *Proc. Int. Conf. on Communications (ICC)*, London, UK, June 2015, pp. 3323–3328.

[15] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, Apr. 2011.

[16] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[17] A. Magesh and V. V. Veeravalli, "Multi-player multi-armed bandits with non-zero rewards on collisions for uncoordinated spectrum access," 2019. [Online]. Available: arXiv:1910.09089

[18] M. Bande, A. Magesh, and V. V. Veeravalli, "Dynamic spectrum access using stochastic multi-user bandits," *to appear in IEEE Wireless Commun. Lett.*, 2021. [Online]. Available: arxiv:2101.04388

[19] S. Ali, A. Ferdowsi, W. Saad, N. Rajatheva, and J. Haapola, "Sleeping multi-armed bandit learning for fast uplink grant allocation in machine type communications," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5072–5086, Aug. 2020.

[20] Y. Lin, T. Wang, and S. Wang, "UAV-Assisted Emergency Communications: An Extended Multi-Armed Bandit Perspective," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 938–941, Mar. 2019.

[21] M. N. Esfahani and B. S. Ghahfarokhi, "Improving spectrum efficiency in fractional allocation of radio resources to self-organized femtocells using learning automata," in *Int. Symp. on Telecommun.*, Tehran, Iran, Sept. 2014, pp. 1071–1076.

[22] J. A. Ayala-Romero, J. J. Alcaraz, and J. Vales-Alonso, "Data-Driven Configuration of Interference Coordination Parameters in HetNets," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 5174–5187, Apr. 2018.

[23] Z. Tian, J. Wang, J. Wang, and J. Song, "Distributed NOMA-based multi-armed bandit approach for channel access in cognitive radio networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1112–1115, Aug. 2019.

[24] M. A. Adjif, O. Habachi, and J. Cances, "Joint channel selection and power control for NOMA: A multi-armed bandit approach," in *Proc. IEEE Wireless Commun. and Networking Conf. (WCNC)*, Apr. 2019, pp. 1–6.

[25] S. Maghsudi and S. Stańczak, "Joint channel selection and power control in infrastructureless wireless networks: A multiplayer multiarmed bandit framework," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4565–4578, Oct. 2015.

[26] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge: Cambridge University Press, 2005.

[27] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Annual Int. Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, Sept. 2013, pp. 611–615.

[28] M. J. Youssef, J. Farah, C. A. Nour, and C. Douillard, "Resource allocation in NOMA systems for centralized and distributed antennas with mixed traffic using matching theory," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 414–428, Jan. 2020.

[29] ——, "Full-duplex and backhaul-constrained UAV-enabled networks using NOMA," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9667–9681, Sept. 2020.

[30] M. J. Youssef, C. A. Nour, J. Farah, and C. Douillard, "Backhaul-constrained resource allocation and 3D placement for UAV-enabled networks," in *Proc. IEEE Veh. Techn. Conf. Fall (VTC)*, Honolulu, Hi, USA, Sept. 2019, pp. 1–7.

[31] I. Bistritz and A. Leshem, "Distributed multi-player bandits - a game of thrones approach," in *32nd Proc. Int. Conf. on Neural Inf. Process. Syst.*, ser. NIPS'18, Montreal, Canada, Dec. 2018, pp. 7222–7232.

[32] J. Choi, "NOMA-Based Random Access With Multichannel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, Dec. 2017.

[33] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

[34] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," in *SIAM J. Control Optim.*, no. 5, Sept. 2014, pp. 2753–2770.
[35] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, 1993. [Online]. Available: http://www.jstor.org/stable/2951778
[36] K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher, "Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified," *29th Symp. Theor. Aspects of Comput. Sci.*, pp. 124–135, Feb. 2012.
[37] 3GPP, "TR25-814 (V7.1.0), Physical Layer Aspects for Evolved Universal Terrestrial Radio Access (UTRA)," 2006.
[38] "5G NR Physical channels and modulation," *3GPP TS 38.211 version 15.3.0 Release 15*, Oct. 2018.
[39] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits-a musical chairs approach," in *Int. Conf. on Mach. Learn.*, 2016, pp. 155–163.
[40] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.

Marie-Josepha Youssef received her B.E. and M.S. degrees in computer and communications engineering in 2016 from the Lebanese University, and her Ph.D. degree in information and communication engineering in 2020 from IMT Atlantique, France. Her current research interests include resource allocation, non-orthogonal multiple access, matching theory, reinforcement learning, unmanned-aerial vehicles and grant-free communications.

Venugopal V. Veeravalli (M'92, SM'98, F'06) received the B.Tech. degree (Silver Medal Honors) from the Indian Institute of Technology, Bombay, in 1985, the M.S. degree from Carnegie Mellon University, Pittsburgh, PA, in 1987, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, in 1992, all in electrical engineering.

He joined the University of Illinois at Urbana-Champaign in 2000, where he is currently the Henry Magnuski Professor in the Department of Electrical and Computer Engineering, and where he is also affiliated with the Department of Statistics, and the Coordinated Science Laboratory. Prior to joining the University of Illinois, he was on faculty of the ECE Department at Cornell University. He served as a Program Director for communications research at the U.S. National Science Foundation from 2003 to 2005. His research interests span the theoretical areas of statistical inference, machine learning, and information theory, with applications to data science, wireless communications, and sensor networks.

He was a Distinguished Lecturer for the IEEE Signal Processing Society during 2010–2011. He has been on the Board of Governors of the IEEE Information Theory Society. He has been an Associate Editor for Detection and Estimation for the IEEE Transactions on Information Theory and for the IEEE Transactions on Wireless Communications. Among the awards he has received for research and teaching are the IEEE Browder J. Thompson Best Paper Award, the Presidential Early Career Award for Scientists and Engineers (PECASE), and the Wald Prize in Sequential Analysis.

Joumana Farah received the B.E. degree in Electrical Engineering from the Lebanese University, in 1998, the M.E. degree in Signal, Image, and Speech processing, in 1999, and the Ph.D. degree in mobile communication systems, in 2002, from the University of Grenoble, France. In 2010, she obtained the Habilitation to Direct Research (HDR) from University Pierre and Marie Curie (Paris VI), France. She is currently a full-time professor at the Faculty of Engineering, Lebanese University, Lebanon. She has supervised a large number of Master, PhD theses and post-docs. She has received several research grants from the Lebanese National Council for Scientific Research, the Franco-Lebanese CEDRE program, and the Lebanese University. She has nine registered patents and a software and has coauthored a research book and over a hundred of papers in international journals and conferences. Her current research interests include resource allocation techniques, channel coding, channel estimation, and interference management techniques. She was the General Chair of the 19th International Conference on Telecommunications (ICT 2012), and serves as a TPC member and a reviewer for several journals and conferences.

Charbel Abdel Nour (Senior Member, IEEE) received the Computer and Communications Engineering degree from Lebanese University, Roumieh, Lebanon, in 2002, the master's degree in digital communications from the University of Valenciennes, Valenciennes, France, in 2003, the Ph.D. degree in digital communications from Telecom Bretagne, Brest, France, in 2008 and the Accreditation to Supervise Research degree from the University of Southern Brittany, Lorient, France, in 2020. From June 2007 till October 2011, he worked as a Post-doctoral Fellow with the Department of Electronics, Telecom Bretagne. He was involved in several research projects related to broadcasting and satellite communications. Additionally during the same period, he was active in the Digital Video Broadcasting DVB Consortium, where he had important contributions. Since November 2011, he holds an Associate Professor position with the Department of Electronics, Telecom Bretagne. Lately, he presented several contributions to the H2020 METIS, FANTASTIC5G and EPIC projects and to the 3GPP consortium related to coding solutions for 5G. His research interests include radio mobile communications systems, broadcasting systems, coded modulations, error correcting codes, resource and power allocation techniques, waveform design, MIMO, and iterative receivers.

Catherine Douillard received the engineering degree in telecommunications from the Ecole Nationale Superieure des Telecommunications de Bretagne, France, in 1988, the Ph.D. degree in electrical engineering from the University of Western Brittany, Brest, France, in 1992, and the accreditation to supervise research from the University of Southern Brittany, Lorient, France, in 2004. She is currently a full Professor in the Mathematical and Electrical Engineering department of IMT Atlantique where she was in charge of the Algorithm-Silicon Interaction research team until 2020. Her main research interests are turbo codes and iterative decoding, iterative detection, the efficient combination of high spectral efficiency modulation and turbo coding schemes, diversity techniques and turbo processing for multi-carrier, multi-antenna and multiple access transmission systems. In 2009, she received the SEE/IEEE Glavieux Award for her contribution to standards and related industrial impact. She was active in the DVB (Digital Video Broadcasting) Technical Modules for the definition of DVB-T2, DVB-NGH as chairperson of the "Coding, Constellations and Interleaving" task force and DVB-RCS NG standards. Since 2015, she has had several contributions in the FANTASTIC-5G and EPIC H2020 European projects intended for the definition of new techniques for 5G and beyond.