Medium Access Control protocol for Collaborative Spectrum Learning in Wireless Networks

Tomer Boyarski and Amir Leshem, Senior Member, IEEE

Abstract—In recent years there is a growing effort to provide learning algorithms for spectrum collaboration. In this paper we present a medium access control protocol which allows spectrum collaboration with minimal regret and high spectral efficiency in highly loaded networks. We present a fully-distributed algorithm for spectrum collaboration in congested ad-hoc networks. The algorithm jointly solves both the channel allocation and access scheduling problems. We prove that the algorithm has an optimal logarithmic regret. Based on the algorithm we provide a medium access control protocol which allows distributed implementation of the algorithm in ad-hoc networks. The protocol utilizes single-channel opportunistic carrier sensing to carry out a low-complexity distributed auction in time and frequency. We also discuss practical implementation issues such as bounded frame size and speed of convergence. Computer simulations comparing the algorithm to state-of-the-art distributed medium access control protocols show the significant advantage of the proposed scheme.

Keywords: Machine learning, Medium Access Control, Spectrum sharing

I. INTRODUCTION

Spectrum collaboration is a fundamental problem in centralized and ad-hoc wireless networks. In centralized networks, the well-known Hungarian Algorithm [1], [2] can be used at the base station to find the optimal orthogonal allocation [3], [4], [5].

In ad-hoc wireless networks no central controller exists to collect the channel state information (CSI) from all users and manage spectrum access.

In the following sections we describe the state-of-the-art approaches to network management and spectrum collaboration. These approaches can be categorized as quasi-centralized (where either a leader is selected or CSI is broadcasted to all participants) and fully distributed approaches where nodes act independently. The fully distributed approaches utilize either ALOHA type medium Access Control (MAC) or Carrier Sensing Multiple Access (CSMA) based MAC.

A. Quasi-centralized algorithms

The simplest approach to solve the spectrum collaboration problem in ad-hoc networks is sharing CSI between nodes, thereby reducing the problem back to its centralized form. This approach is taken in [6] and [7] for stationary channels and in [8], [9] and [10] for dynamic channels.

Tibrewal et al. [6] used carrier-sensing. This facilitated the gathering of all CSI in every node where the Hungarian Algorithm was applied to find the optimal allocation. In [7] no carrier sensing capabilities were assumed. Therefore, signaling between users become more costly. Hence, rather than collecting the CSI in every node like in [6], in [7] the CSI was gathered only in a single Leader node.

Rather than using the Hungarian Algorithm as in [6], the Leader node in [7] applies a matching elimination algorithm and signals its results to all other users. The matching elimination algorithm proceeds by eliminating edges from the bipartite graph representing the resource allocation problem. Edges that are not included in high-welfare matchings are eliminated from consideration. The Followers improve their estimation of the surviving edges until each user only has one surviving edge and the algorithm proceeds to the exploitation phase.

In [8] a single node applies the Exponential-weight algorithm for Exploration and Exploitation (EXP3) of [11] with multiple plays. The optimal regret for dynamic channels is $O(\sqrt{T})$ while the regret of [8] is $O(T^{2/3})$. By restricting their discussion to a two-users scenario, the authors of [9] were able to achieve optimal regret. The algorithm in [9] relied on a filtering strategy with low switching and was inspired in part by the Shrinking Dartboard algorithm [12].

Another variant of EXP3 with direct messaging between users achieved optimal regret for the multi-user scenario over dynamic channels in [13]. This was possible because the adversary was assumed to be oblivious and the utilities were not user-dependent. This version was later improved in [10] such that optimal regret was achieved for each user individually.

Since these approaches emulate a centralized decision they can achieve good results. For example the First and third place winners of the recent DARPA spectrum collaboration challenge(teams GatorWings [14] and Zylinium [15]) used a leader election based control. GatorWings modeled the problem as a Partially Observable Markov Decision Process (POMDP) and implemented a SARSA algorithm using a Deep Neural Network (DNN). Zylinium modeled the problem as a mixed integer programming problem and solved it with Google’s OR-Tools library [16]. While in theory the regrets of [6], [7], [8], [9] and [10] are attractively close to their respective lower bounds, in practice the signaling of CSI results in significant overhead and these methods do not scale well. Furthermore, leader election is non-robust especially in mobile ad-hoc networks - if the leader is jammed or becomes disconnected due to mobility. These problems, suggest that a more robust fully distributed solution might be the way to
achieve robust spectrum sharing. In the following subsection we discuss two families of solutions where random access protocols are used to obtain an optimal orthogonal allocation.

B. Fully distributed random access based spectrum collaboration

An alternative to these quasi-centralized algorithm are contention-based methods such as Aloha and CSMA. Unfortunately, traditional Aloha and CSMA achieve less than 80% spectrum utilization in most realistic scenarios, and their performance is especially bad in loaded networks [17]. Therefore, to approach optimal utilization, approaches based on random access are used to converge to an optimal orthogonal allocation. Examples for such approaches utilizing Aloha are [18], [19], [20] and [21]. Similarly, CSMA based solution have also been suggested, e.g. in [22], [23], [24], [25], [26], [27].

These variants aim to achieve orthogonal allocations in a fully-distributed manner without any information sharing between nodes. Some of these variants aim to achieve optimal OFDMA allocations where the sum of the Quality of Service (QoS) of all the users is maximized [18], [19], [20], [25], [26], [27] while others aim to achieve stable OFDMA allocations where no user can unilaterally improve its QoS [22], [23], [24], [29]. Recently, new approaches incorporating fairness have also been proposed [21],[30].

1) Aloha based methods: The distributed spectrum collaboration problem in wireless ad-hoc networks can be modeled as a multi-arm bandit problem under two conditions: No carrier-sensing capabilities are assumed, but collision data can be sent to the transmitter and the number of channels is equal to or greater than the number of users. The simpler formulation is that of a single-user Multi-Armed Bandit (MAB) problem. The more comprehensive formulation is that of a multi-user MAB problem (MUMAB).

For the single user case the problem is equivalent to that of a Restless MAB problem. In this model each channel follows its own Markov chain with an unknown transition matrix and the states are observable to the users. When the channels and users are homogeneous, this can be solved with Deterministic Sequencing of Exploration and Exploitation without any explicit signaling between users [31]. Under the same homogeneity assumption, the problem can also be solved with a Partially Observable Markovian Decision Process [32] or with Whittle’s index policy [33]. Another variant of the MAB formulation is the Hidden Markov Bandit. In this model all channels follow a shared Markov chain with an unknown transition matrix and the state is not observable to the users. Logarithmic regret was achieved for this formulation in [34].

Generalizations to multi-user competitive MDP’s has recently been proposed using multi-agent Deep Q-Learning (DQL) [35]. In this algorithm the users collect data on their environment constantly and locally. However, their decisions are based on a model that is trained and updated periodically and globally in a central unit (e.g. cloud). Deep Q-learning was also used in [36] only considered channels with high Q-values. From those, the least-loaded channel was selected. The authors proved that this method is asymptotically optimal for large overcrowded networks using insights from game theory and combinatorial optimization. The method was shown to outperform [33] over stationary channels with computer simulations.

In [37] DRL is used to maximize the total throughput or fairness when neighboring networks are using unknown MAC protocols such Aloha or TDMA. In [29] the authors achieve a stable allocation for the Restless MUMAB problem with heterogeneous channels. In [38] DQL for Markovian channels are considered where $\varepsilon$-greedy policies are used to accumulate the training data. DRL for power control of secondary users is explored in [39]. In [40] the authors apply DRL for efficient co-existence with time division multiple access (TDMA) and ALOHA nodes.

All these techniques can be proven to be optimal from a single user perspective, when ignoring the actions of the other players or treating them as random. However, since all users act selfishly, this becomes a Markov game, and the solution is a Nash equilibrium of the corresponding Markov game. It is well known that such Nash Equilibrium points can be sub-optimal. Indeed, Multi-agent reinforcement learning converges to some Nash Equilibrium (NE) of the game. This NE may be highly inefficient [41]. Furthermore, these reinforcement learning methods require passing information between users which causes significant overhead or pre-training. To overcome the inefficiency caused by competitive solutions, the distributed spectrum collaboration problem in wireless ad-hoc networks can also be formulated as a collaborative Stochastic Multi-User multi-arm bandit problem. A significant advantage of this formulation is that it does not require passing information between users. An Aloha-style collaborative game for this formulation was presented in [18] and had a regret bound of $O(\ln^{2+\delta} T)$ for some small $0 < \delta < 1$. The authors refined their algorithm in [19] and reduced the regret bound to a near-optimal $O(\ln^{1+\delta} T)$ for any $\delta > 0$. The algorithms in [18] and [19] were restricted to non-congested networks where the number of channels was greater than the number of users. This algorithm was inspired by a Perturbed Markov Chain Approach from [42] which solves the distributed optimization for a known game. A related approach was presented in [20] for IoT networks where contextual bandit model has been proposed. Min-Max fairness was achieved in [21], [30] while maintaining a near-optimal regret bound of $O(\ln T \ln \ln T)$.

2) Carrier sensing based methods: Whereas collision based multi-access protocols such as Aloha and its advanced variants [19], [20] are appropriate for simple devices such as sensor networks they are inefficient for wireless communication ad-hoc network since convergence time to the optimum can be significant. In particular, the convergence time of Markov chain based method may be exponential in the number of users and channels. In practice, most random access networks utilize a version of the listen before talk which is the basis to CSMA.

Opportunistic CSMA [43] with collision avoidance was first used in [22] to distributedly find a stable configuration (one-to-one matching) with an enhanced version of the Gale-Shapley algorithm. This approach was later extended for the case where several resources can be assigned to a single user
(many-to-one) in [45] where the same resource can be assigned to different users simultaneously (many-to-many matching). The many-to-one scenario was revisited in [46] with improved reliability and latency. These improvements were achieved by utilizing the optimal connectivity approach for each user, optimizing the maximum number of matched resources, and providing a resource reservation mechanism for users suffering from bad channel conditions. Stable Marriage with cheating was introduced in [47] and with social awareness in [48]. Swap-based algorithms were also developed for distributed learning of stable allocations [23], [24]. These algorithms have versions where users are allowed to leave and enter the network and their probability not to be in a stable orthogonal allocation decreases exponentially with time. In these algorithms users initiate and accept swaps through signaling until no user wishes to change its resource. Assuming the role of swap-initiator can be done in a round-robin fashion as in [23] or a dynamic fashion as in [24]. Assumptions regarding carrier-sensing capabilities also vary among these algorithms. Some algorithms assume wide-band sensing [23] while others only single-band sensing [23]. A common feature of these algorithms is the use of the upper confidence bound to determine channel preferences. Unfortunately, since a stable allocation is not necessarily optimal, no regret guarantees can be made regarding these algorithms.

Recently, OCSMA was used in combination with the well-known auction algorithm [49] to find the optimal allocation via collaborative games in non-congested networks where the number of channels is greater than the number of users [25], [26], [27], [28]. In [28], authors used the original centralized auction algorithm [49]. In this algorithm, the users collect their CSI locally and then transmit their bids to each other. Unfortunately, this approach incurs a significant overhead.

To avoid excessive overhead, a distributed version of the auction algorithm [26] was proposed in [27]. Unfortunately, the contention window in [27] grows exponentially with the number of collisions and the number of auction iterations until convergence may also be quite large.

To reduce the number of distributed auction iterations until convergence, the authors of [28] proposed a binary utility matrix (in the extended context of learning energy efficient channel collaboration). In this matrix a channel is considered "good" if it is one of the $O(\ln(N))$ best channels and "bad" otherwise. Unfortunately, this method does not guarantee optimality. Therefore, it may suffer linear regret.

**C. Contribution:**

In this paper we consider congested ad-hoc wireless networks where the number of users can be much larger than the number of channels. For such networks, we present a full MAC protocol that learns the jointly optimal spectrum allocation and scheduling. In contrast to previous works, e.g. [27], [26], [25] we allow more users than channels, and solve this, by a protocol converging to the optimal FDMA/TDMA allocation with logarithmic regret, which means that the learning overhead decreases as $O(\ln^2 T)$. Additionally, we provide finite block length implementation, and bound the overhead for this case, by relaxing the theoretically optimal results, thus providing full understanding of the trade-off between exploration and exploitation, in fixed block length implementation.

This is a significant step towards practical implementation of collaborative optimal distributed management of ad-hoc networks, since practical systems, require fixed framing structure. This is the first fully-distributed algorithm for this scenario with a provable optimal regret of $O(\ln T)$. Moreover, our analysis provides a bounded overhead protocol with fixed frame size, where the overhead can be arbitrarily low, and the probability of bad allocation decreases exponentially with time. This follows from the theoretical analysis of the exponentially growing exploitation window algorithm. The significant advantage of the fixed frame size is the scalability on one hand, and the possibility to adapt the allocations upon network changes.

On the computational side, we significantly accelerate the auction algorithm using $\varepsilon$-scaling to reduce the number of distributed auction iterations until convergence. This acceleration method is superior to that of [28] since it guarantees optimality.

Furthermore, we develop a collision-resolution mechanism to cope with finite size of the contention window. This method is superior to that of [27] where the contention window size is not bounded. Our collision-resolution mechanism only adds a factor of $O(\ln N)$ to the total regret. We demonstrate the high efficiency of our protocol using simulation over 5G channels.

The rest of this paper is organized as follows. In section II we describe the resource sharing problem. In section III we describe our MAC protocol and the algorithms used to implement the protocol. In section IV we analyze the expected regret of the MAC, under the assumption of static network and exponentially growing exploitation phase. In section V we discuss the implementation of the protocol in dynamic environments where exponentially growing exploitation is undesirable. We further suggest that during the first epoch, the exploration phase should be longer than the subsequent epochs to allow good starting point for the learning and adaptation. In section VI we provide simulations. First we demonstrate the logarithmic regret under stationary channel and exponentially growing exploitation and then show in simulations, that limiting the exploitation phase to be a large and significant part of the the frame leads to excellent utilization of the spectral resources even for dynamically changing networks. We end with some concluding remarks.

**II. System Model**

Consider an ad-hoc wireless network with $N$ learning user pairs with carrier sensing capabilities. A schematic depiction of our network is presented in Figure 1. These users share $K$ frequency channels. Our network is congested, i.e., $N \gg K$ so time division is necessary. Time/frequency slots are called resources. The users apply CSMA only during learning phases, while data transmission phase uses orthogonal allocation of resources.

For simplicity we assume that the average congestion on each channel $M \triangleq N/K$ is an integer and that each user
transmission rate is a discrete quantity with some resolution \( \text{MCS} \). This translates to a finite set of data rates, i.e., small number of possible coding and modulation schemes.

\[ a_n \triangleq (k_n, m_n) \]

A OFDMA configuration with 12 users and 3 channels is depicted in Figure 2. Since users select resources distributedly, we refer to such a selection as an action. The action of user \( n \) is a tuple of channel and time slot choices \( a_n \triangleq (k_n, m_n) \) where \( k_n \in [K] \) and \( m_n \in [M] \) and \( |M| = \{1, \ldots, M\} \). action profile is a vector of actions by all users \( a \triangleq (a_1, \ldots, a_N) \).

Finally, we assume that each users pair can listen and transmit simultaneously they block each other and their respective rates are close to zero. For dynamic channels, \( Q_{n,a} \) is time varying and also different time slots at the same frequency might suffer different interference which might also be time varying, therefore, we use \( a \) as the QoS index. Since users apply coding for the fading channel, long term averages of \( Q_{n,a} \) define the performance. Unfortunately, these are not known to the users a-priori and each user needs to estimate these parameters by sampling the channels repeatedly and estimating the QoS. The QoS sample of user \( n \)'s action at time \( t \) is denoted by \( q_{n,a}^t \) and we assume that \( \forall t : E(q_{n,a}^t) = Q_{n,a} \). Furthermore, \( q_{n,a}^t \) is bounded between \( Q_M \) and \( Q_{\min} \) and i.i.d. in time. Let \( W(\mathbf{a}) \) be the social welfare of a given action profile \( \mathbf{a} \) defined by:

\[ W(\mathbf{a}) \triangleq \sum_{n=1}^N U(n, \mathbf{a}). \quad (3) \]

Let the optimal action profile \( \mathbf{a}^* \) be defined by

\[ \mathbf{a}^* = \arg \max_\mathbf{a} W(\mathbf{a}), \quad (4) \]

and let the optimal social welfare

\[ W^* \triangleq W(\mathbf{a}^*). \quad (5) \]

Let \( Q = [Q_{n,a}] \) be the utility matrix of the users. Time is slotted and indexed by \( t \) and that the users are active for \( T \) time slots. \( T \) and \( Q \) are not known to the users and there is no explicit message passing between users, except a feedback channel between each transmitter and its designated receiver. Therefore, \( \mathbf{a}^* \) must be learned in a fully-distributed manner. The welfare at time \( t \) is \( W^t \). The regret is defined with respect to \( W^* \) by:

\[ R \triangleq E \left[ TW^* - \sum_{t=1}^T W^t \right] \quad (6) \]

where the expectation is taken with respect to the randomness of the utilities and the users’ choices. In order to minimize the regret the users must learn \( \mathbf{a}^* \) as quickly as possible. To derive a logarithmic regret users must be able to estimate \( a^* \) with an exponentially decaying probability of error.

As we will show in our main theorem regret of \( O(\ln T) \) is achievable. This is very satisfying, since the relative management overhead reduces to 0 quickly at rate \( O(\frac{\ln T}{T}) \).

Our goal in this paper is to develop a fully distributed MAC protocol which allows the users to independently and distributively learn an optimal action profile \( \mathbf{a}^* \) with logarithmic regret. We also study sub-optimal, fixed block length versions of the protocol and bound their regret.
III. MEDIUM ACCESS CONTROL PROTOCOL FOR COLLABORATIVE SPECTRUM ACCESS

In this section we describe a medium access control protocol for ad-hoc wireless networks combined with a learning algorithm for achieving a regret optimal resource allocation. The protocol is divided into epochs and each epoch consists of two phases: Learning and exploitation phases; these phases manifests the classical trade-off between exploration and exploitation. Indeed, properly determining the length of each phase is crucial for achieving optimal performance. The learning phase is further divided into two sub-phases: Exploration phase, coordination phase. The first is used to learn the channel state information for each user independently, while the coordination phase is used to obtain an optimal (with respect to the estimated channel state information) time-frequency allocation \( \hat{a}' \).

During the exploitation phase, each user transmits its data using the estimated optimal allocation \( \hat{a}' \). Epochs are indexed by \( j \), and the duration of the coordination and exploitation phases during the \( j \)’th epoch are denoted by \( T_j^1, T_j^2 \) and \( T_j^3 \), respectively. We will show that by setting \( T_j^1 = T_j^2 = T_2^j \) to be independent of \( j \) while \( T_j^3 = 2^j \) grows exponentially, leads to a logarithmic regret. In the rest of this section we discuss the structure of each of the phases the frame structure and the state machine of the protocol.

The structure of the algorithm is depicted in Figure 3. The main pseudo-code of the algorithm appears in algorithm 1. We now describe the learning phases of the algorithm in detail.

These phases occur in the beginning of each epoch.

A. Exploration Phase

The first learning phase is the exploration phase. In the beginning of the phase the user initializes its utility sample sum on each resource \( a = (k, m) \) to be zero \( S_{n,a} = 0 \) and its visitations counter \( V_{n,a} = 0 \). In each time slot of this phase the user chooses a channel and time slot uniformly at random and transmits a pilot signal. If the signal is properly received an ACK message containing the estimated utility is sent by the receiver. If ACK is received for this transmission, the transmitter updates the Utility Sample Sum \( S_{n,a} \) and the Visitations Counter \( V_{n,a} \) according to eq. sample sum and its visitations counter:

\[
V_{n,a} \leftarrow V_{n,a} + 1, \quad S_{n,a} \leftarrow S_{n,a} + u_{n,a} \tag{7}
\]

At the end of this phase each user estimates its QoS for each action by

\[
\hat{Q}_{n,a} \leftarrow \frac{S_{n,a}}{V_{n,a}} \tag{8}
\]

Afterwards, each user adds a small dithering to each estimated QoS to make the optimal action profile for the estimated matrix unique with probability 1. The maximal dithering is \( d_{\text{max}} \approx \frac{\Delta_{\text{max}}}{8N} \). In Appendix A, Lemma [11.4], we will prove that the dithering does not change the optimal action profile. The dithering of the utility of user \( n \) on channel \( k \) and slot \( m \) is uniformly distributed \( D_{n,a}^m \sim U([-d_{\text{max}}, d_{\text{max}}]) \) and i.i.d. in \( k \) and \( m \). The dithered estimated utilities are:

\[
\hat{Q}_{n,m} \leftarrow \hat{Q}_{n,a} + D_{n,a} \tag{9}
\]

The pseudo-code of the exploration phase appears in algorithm 2.

B. Auction Phase

In the second phase the users hold a distributed auction over the resources. In the beginning of this auction all users are unassigned. Users become assigned by bidding over resources and winning them in iterations. The maximal number of iteration is \( I_{\text{max}} \). At the end of each iteration there is an unassigned notification slot. In this slot all receivers of unassigned transmitters transmit on channel 1. This modest feedback incurs very little overhead. If a transmission occurred the auction continues. Otherwise, it terminates. The structure of the auction phase is depicted in Figure 4. The main pseudo-code of the auction phase appears in algorithm 3.

Each iteration is composed of \( M \) frames. Bidding over resource \( a = (k, m) \) is done over channel \( k \) in frame \( m \). This is depicted in figure 5. Each frame has a deterministic part and a random part. The purpose of these parts is to cope with the finite resolution of the contention window, i.e. what to do when bids of two users over the same resource fall into the same CSMA slot.

In the beginning of the auction the user \( n \) initializes all entries of its bid matrix to be zero \( B_n = 0 \). The minimal bid increment is \( \varepsilon \). It is initialized to some value \( \varepsilon_0 \) (e.g. \( \varepsilon_0 = \Delta_{\text{min}} \)) and multiplied by a scaling factor \( 0 < \zeta < 1 \) in each iteration until it reaches its final value \( \varepsilon^* \approx \frac{\Delta_{\text{min}}}{8N} \). Let \( \varepsilon^i \) be the \( \varepsilon \) of auction iteration \( i \). The \( \varepsilon \)-scaling update rule is:

\[
\varepsilon^{i+1} \leftarrow \max(\varepsilon^*, \varepsilon^i \zeta) \tag{10}
\]

Users bid over their most profitable resource. The profit of user \( n \) with resource \( a \) is the difference between its estimated QoS and its bid:

\[
g_{n,a} \leftarrow \hat{Q}_{n,a} - B_{n,a} \tag{11}
\]

The most-profitable resource of user \( n \) is:

\[
a^*_n \triangleq \arg \max_a (g_{n,a}) \tag{12}
\]

Re-order the profits of user \( n \):

\[
g_{n,(1)} \leq g_{n,(2)} \leq \ldots \leq g_{n,(N-1)} \leq g_{n,(N)} \tag{13}
\]

Let the profit gap \( g_{n}^{\text{gap}} \) of user \( n \) be:

\[
g_{n}^{\text{gap}} \triangleq g_{n,(N)} - g_{n,(N-1)} \tag{14}
\]

The bid increment of user \( n \) is:

\[
B_n^\Delta = \varepsilon + g_{n}^{\text{gap}} \tag{15}
\]

The new bid of user \( n \) on resource \( a \) is:

\[
B_{n,a} \leftarrow B_{n,a} + B_n^\Delta \tag{16}
\]

The pseudo-code of the update of the continuous bid appears in algorithm 4.
We now begin the process of translating the continuous bid into a discrete bid. First, note that the maximal discrete bid which suffices for the distributed auction algorithm $b^*$ is given by:

$$b^* = 8NQ_M/\Delta_{\text{min}}.$$  
(17)

In practice optimizing $b^*$ can improve the performance of the MAC. Let the relative continuous bid of user $n$ on resource $a$ be $B_{n,a}/Q_M$. Let

$$\rho_{n,a} = 1 - B_{n,a}/Q_M.$$  
(18)

Let $\beta$ be the number of slots in each contention window (auction block). Let us consider the representation of the relative complementary continuous bid of user $n$ on resource $a$ in base $\beta$:

$$[\rho_{n,a}]_{\beta} = 0, \rho_1 \cdots \rho_\lambda, \cdots$$  
(19)

where $\lambda$ represents the required resolution of the bids in the distributed auction algorithm:

$$\lambda \triangleq \log_\beta b^*.$$  
(20)

Each frame is composed of at most $\lambda$ deterministic blocks each containing $\beta$ slots followed by a collision notification slot as depicted in Figures 6 and possibly several random collision resolution blocks with two slots each. The role of the deterministic blocks is to identify the highest bidders (up to accuracy $Q_M/b^*$) and then the random collision resolution blocks select one of these highest bidders.

Each user initializes its local state to undetermined. In deterministic block $i$ user $n$ begins to transmit on slot $\rho_i+1$ as defined in (19). If any user began transmitting before user $n$ then user $n$ sets its state to loser. After each deterministic block there is a collision notification slot where a NACK message is transmitted by each receiver which failed to decode the message properly. If no transmission occurs in this slot then the single remaining undetermined user changes its state to winner and the bidding over this resource in this iteration is finished. If there is a transmission in this slot this means that a more refined resolution is necessary to determine the winner and the algorithm proceeds either to another deterministic block (in the first $\lambda$ blocks) or to a random block if resolution is sufficiently high as explained in Lemma IV.10.

The over all collision-resolution flow (both deterministic and random) is depicted in Figure 8. A pseudo-code describing a single auction iteration with its frames and blocks appears in algorithm 5.

### IV. REGRET ANALYSIS

In this section we assume that the exploitation phase grows exponentially and that the network state is static and prove a logarithmic regret upper bound. To that end we make the following assumption regarding the phases:

1. We found that $\beta = 4$ leads to good performance.

Figure 3: The duration of the exploration, auction, and exploitation phases of epoch $j$ are denoted $T_1^j, T_2^j$ and $T_3^j$, respectively.

Figure 4: There are at most $I_{\text{max}}$ iterations in the auction phase. Unassigned users transmit on channel 1 in the unassigned notification slot.

Figure 5: Each auction iteration is composed of $M$ frames. Users bid over resource $a$ in channel $k$ and frame $m$. 

**Exploration Phase:** In the first phase, the users choose channels uniformly at random for

$$T_1 = 10N^3Q_M^2\Delta_{\text{min}}^{-2}$$  
(21)

slots and estimate their QoS. Users who choose the same channel at the same time do not obtain any estimation. This phase follows the discussion in subsection III-A.

**Auction Phase:** In the second phase the users perform the distributed auction algorithm for

$$T_2 = cN^3\ln N$$  
(22)

slots for some constant $c$ as described in detail in subsection III-B.

**Exploitation Phase:** In the third phase the users exploit the allocation obtained at the end of the auction phase for

$$T_3^j = O(2^j).$$  
(23)

Because the duration of the learning phases is constant and the duration of the exploitation phase increases exponentially with the epoch index, the regret (overhead) of the algorithm...
Lemma IV.1. The estimated utility of user \( n \) with action profile \( a \) is:

\[ U^j_n(a) \triangleq Q^j_{n,a,n} 1 \quad (\forall n : a_n \neq a_n) \]  

The total estimated welfare \( \hat{W}^j(a) \) and the estimated optimal action at epoch \( j \) are respectively defined by:

\[ \hat{W}^j(a) \triangleq \sum_{n=1}^N U^j_n(a), \quad \hat{a}^* \overset{\text{def}}{=} \arg\max_a \hat{W}^j(a) \]  

We now present two lemmas that are necessary for the proof of the main theorem:

Lemma IV.2. The estimated optimal action profile is incorrect with probability \( P(\hat{a}^* \neq a^*) = O(2^{-j}). \)

Lemma IV.3. If \( \varepsilon^* < \frac{\Delta_{\min}}{8N} (1 + \frac{1}{10N}) \) then the auction converges to the estimated optimal allocation \( \hat{a}^* \) within \( \frac{N^3 Q_M}{\Delta_{\min}} \) iterations.

In subsection IV-A we analyze the probability of failure due to the exploration phase and prove lemma IV.2. In subsection IV-B we analyze the probability of failure of the Auction phase and prove lemma IV.3.

A. Exploration Phase Analysis

We now analyze the performance of the exploration phase and its contribution to the accumulated regret.

Definition IV.2. Define the QoS estimation errors at the end of the exploration phase of epoch \( j \):

\[ \xi^j_{n,a} \triangleq \left| \hat{Q}^j_{n,a,n} - Q^j_{n,a} \right|, \quad \xi^j \triangleq \max_{n,a} \xi^j_{n,a} \]  

Let \( \xi_{\max} \) be defined by:

\[ \xi_{\max} \triangleq \frac{3\Delta_{\min}}{8N} \]  

Let \( E^j \) be the error event that \( \hat{a}^* \neq a^* \).

Lemma IV.4. We now prove that as long as \( \xi \leq \xi_{\max} \) the correct action profile is selected, i.e., \( P(E^j) \leq P(\xi \leq \xi_{\max}) \).

The proof is given in appendix A.

Let the number of visitation to the least-visited channel at the end of the exploration phase of epoch \( j \) be \( V^j \triangleq \min_{n,a} V^j_{n,a} \).

Let \( V^j_{\min} \) be defined by:

\[ V^j_{\min} \triangleq \frac{5jN^2}{2\Delta^2}, \]  

where \( \Delta \) is defined in (1).

Let \( E^j_{\min} \) be the event where we fail to obtain at least \( V^j_{\min} \) samples of some resource.
Algorithm 3 Auction - Run in parallel for each user $n$

1: Input: $\hat{Q}_n$, $\Delta_{\min}$, $b^*$, $\zeta$, $Q_M$.
2: Set the initial and final minimal continuous bid increments to $\varepsilon = \Delta_{\min}/4$ and $\varepsilon^* = \Delta_{\min}/8N$.
3: Set the local-user-state to unassigned.
4: Set all the continuous bids to zero: $\forall a : B_{n,a} = 0$.
5: Set global state all-assigned to false.
6: while not all-assigned do
7: $[B_n, a_n] \leftarrow \text{Calculate_Bid} \left( B_n, \hat{Q}_n, \varepsilon \right)$ according to algorithm 2.
8: Perform $\varepsilon$-scaling according to eq. (10).
9: Get discrete bid $b_{n,a_n}$ according to eq. (19).
10: $[a_n, \text{local-user-state}] \leftarrow \text{Place_Bid}(a_n, b_{n,a_n})$ according to algorithm 5.
11: All unassigned users transmit on the all-assigned signalling slot. If no transmission occurred in that slot set global state all-assigned to true.
12: end while

Algorithm 4 Calculate_Bid

1: Input: $B_n$, $Q_n$, $\varepsilon$.
2: if state = assigned then
3: Choose the same resource and keep the same bid.
4: else
5: Calculate profits according to eq. (11).
6: Choose best resource according to eq. (12).
7: Find best and second-best profits according to eq. (13).
8: Find profit gap according to eq. (14).
9: Determine bid increment according to eq. (15).
10: Increase bid according to eq. (16).
11: end if

Lemma IV.5. $P \left( E_V^j \right) = o(2^{-j}).$

Proof: Let $P_{ss}$ be the probability that a user will successfully sample a channel in an exploration slot. This probability is given by:

$$P_{ss} = \frac{1}{N^N} \left( 1 - \frac{1}{N} \right)^{N-1}$$  \hspace{1cm} (29)

With the help of the union bound and Hoeffding’s inequality for Bernoulli’s distribution we obtain:

$$P \left( E_V^j \right) \leq 2N^2 \exp \left( -20jN^3 \Delta^{-2} \left( \frac{1}{4K} - P_{ss} \right)^2 \right)$$  \hspace{1cm} (30)

Details of the computation are given in appendix B.

Lemma IV.6. $P \left( E_V^j | \sim E_V^j \right) = o(2^{-j}).$

Proof: We rely on lemma IV.4 Hoeffding’s inequality for bounded random variables together with a union bound on the users and channels results in

$$P \left( E_V^j | \sim E_V^j \right) \leq 2NK \exp \left( -2\varepsilon^2 \max \frac{V_{min}^j}{Q_M^2} \right)$$  \hspace{1cm} (31)

Substituting $V_{min}^j$, $\xi_{max}$ and $\Delta$ into (31), we obtain:

$$P \left( E_V^j | \sim E_V^j \right) \leq 2N^2 \exp \left( -\frac{90j}{128} \right)$$  \hspace{1cm} (32)

Algorithm 5 Place_Bid

1: Input: $a_n, b_n$.
2: Set local-user-state to undetermined.
3: Set global-state all-determined to false.
4: Initialize block index to 1.
5: while global-state all-determined is false do
6: if local state = undetermined then
7: if block index $\leq \lambda$ then
8: Choose slot according to equation (19).
9: else
10: Choose slot uniformly at random.
11: end if
12: Begin transmitting on the chosen slot.
13: if Someone else transmitted before I did then
14: Set state to Loser.
15: end if
16: if No transmission in collision notification slot then
17: Set local state to Winner.
18: Set global state all-determined to true.
19: end if
20: end if
21: Increase block index by 1.
22: end while

Details of the bound computation are given in appendix C.

Corollary IV.6.1. $P \left( E_V^j \right) = o(2^{-j}).$

Proof: Clearly $P \left( E_V^j \right) \leq P \left( E_V^j \right) + P \left( E_{\sim V}^j \right)$. Lemmas IV.5 and IV.6 and $\frac{N}{\Delta} > 3$ complete the proof.

B. Auction Phase Analysis

We now analyze the performance of the auction phase and its contribution to the accumulated regret. We discuss the convergence time of the auction in lemmas IV.7 and IV.8 and the performance upon convergence in lemma IV.10.

Lemma IV.7. The auction converges within at most $\frac{N}{\Delta_{\min}} (1 + \frac{1}{10N})$ iterations.

Proof: Lemma 3 of [26] guarantees that user $n$ will become assigned within $I_n \leq KM + \frac{1}{\varepsilon^*} \sum_{a=1}^{N} \hat{Q}_{n,a}$ iterations. A simple and crude union bound ensures that all users will become assigned within $I \leq N \Delta_{\min}$ iterations. Since $\hat{Q}_{n,a} \leq Q_M + \varepsilon \Delta$, and $\varepsilon \Delta = \Delta_{\min}$, the result follows.

We would like to emphasize that the union bounds used are very crude. In practice we know based on the analysis of the assignment problem by Aldous [50] that at about half the users are allocated their best channel, quarter obtain their second best, an so on, so that convergence of the auction is typically much faster. Indeed in our simulations, we limited the auction phase to be less than 1000 iterations, which provided little penalty in efficiency.

We now analyze the effect of the collision-resolution process on the expected number of blocks in each iteration.
Lemma IV.8. The expected number of collision-resolution blocks per frame (both deterministic and random) is upper bounded by $O(\ln N)$ and the winner has the best bid up to a resolution of $Q_M/b^*$. 

Proof: 
The number of deterministic collision resolution blocks is bounded by $\lambda = O(\ln N)$. To finish the proof we need to prove that the expected number of random collisions is $o(\ln N)$. Our analysis will actually show that it is $O(\ln \ln N)$.

Consider what happens when $\ell_a$ users offer the same discrete bid for resource $a$ in the same iteration. Out of these $\ell_a$ users, $\ell_a^1$ users chose the first slot. When $\ell_a^1 = 0$, all of them continue to the next round. Therefore, the expected number of users in the next block is:

$$
E(\hat{N}_{a}^B) \leq \max_n (1, C_1 \ln \ell_a)
$$

(34)

The number of users who choose a resource $a$ in a given iteration is a binomial random variable $Z_a \sim B(N, 1/\lambda)$. Naturally, $\ell_a \leq Z_a$. According to [51] corollary A.1.10, we have:

$$
P(\ell_a \geq x) < \exp(x - (x + 1) \ln(1 + x))
$$

(35)

Clearly, $\exp(x - (x + 1) \ln(1 + x)) < \left(\frac{x}{2}\right)^x$. Set $x = \ln^2(N)$. For $N \geq 12$ we have $\left(\frac{x}{2}\right)^x < N^{-2}$. From eq. (35) we have:

$$
P(\ell_a \geq \ln^2 N) < N^{-2}
$$

(36)

From (34) and (36) the expected number of blocks for TDMA slot $m$ and channel $k$ is upper bounded by:

$$
E(\hat{N}_{a}^B) \leq E(N_{a}^B|\ell_a \geq \ln^2 N) + E(N_{a}^B|\ell_a < \ln^2 N) \leq N^{-2}C_1 \ln N + C_1 \ln (\ln^2 N)
$$

(37)

Recall that bidding over different frequencies occurs simultaneously. Therefore, the expected number of blocks required in frame $m$ is $\hat{N}_{a}^B \triangleq \max_k N_{a}^B$. The probability that $N_{a}^B$ is greater than $C_1 \ln \ln^2 N$ is upper bounded as follows:

$$
P(N_{a}^B > C_1 \ln \ln^2 N) = 1 - P(N_{a}^B < C_1 \ln \ln^2 N) = 1 - \left[P(N_{a}^B < C_1 \ln \ln^2 N)\right]^K = 1 - \left[1 - N^{-2}\right]^K = \frac{K}{N^2}
$$

(38)

Therefore, the expected number of random blocks per frame is upper bounded as follows:

$$
E[\hat{N}_{a}^B] \leq 2C_1 \ln N + \frac{K}{N^2}C_1 \ln N.
$$

(39)

Lemma IV.9. If $a$ and $a'$ are two action profiles such that $|\hat{W}(a) - \hat{W}(a')| \neq 0$ then $|\hat{W}(a) - \hat{W}(a')| \geq \frac{\Delta^\text{min}}{2}$.

The proof appears in appendix D.

Lemma IV.10. The distributed auction algorithm [24] converges to $\hat{a}^*$ as long as $\varepsilon^* \leq \frac{\Delta^\text{min}/8}{N}$.

Proof: 
If user $n$ is assigned to resource $a$ then local $\varepsilon$-complementary slackness guarantees:

$$
\hat{Q}_{n,a} - B_{n,a} \geq \max_m \left(\hat{Q}_{n,m} - B_{n,m}\right) - \varepsilon
$$

(40)

Let the maximal bid on resource $a$ be:

$$
\mathbb{E}_a \triangleq \max_n B_{n,a}
$$

(41)

Since $\mathbb{E}_a \geq B_{n,a}$ $\forall n$, we have:

$$
\max_a \left(\hat{Q}_{n,a} - B_{n,a}\right) \geq \max_a \left(\hat{Q}_{n,a} - \mathbb{E}_a\right)
$$

(42)

Let $L$ be the precision of the quantization function that maps the continuous bids to discrete bids. For the quantization function of equation (19) we have $L = Q_M/b^*$. Since resource $a$ is assigned to user $n$ we have:

$$
B_{n,a} > \mathbb{E}_a - L
$$

(43)

Setting (43) into the left-hand side of (40) and setting (42) into the right-hand side of (40) we obtain:

$$
\hat{Q}_{n,a} - \mathbb{E}_a \geq \max_a \left(\hat{Q}_{n,a} - \mathbb{E}_a\right) - (\varepsilon + L)
$$

(44)

Let $\gamma \triangleq \varepsilon + L$. Equation (44) is the global $\gamma$-Complementary Slackness of the original auction algorithm of [19]. Let the quantization function be that of equation (19). By the end of the deterministic collision resolution phase, the winner is within $L = \frac{\Delta^\text{min}}{8N}$ from the highest bidder (since after that step, the collision resolution is random). Coupled with $\varepsilon = \frac{\Delta^\text{min}}{8N}$ we obtain $\gamma = \frac{\Delta^\text{min}}{4N}$. Let $\hat{a}^*$ be the estimated optimal action profile of eq. (25). Let $\hat{a}$ be the auction allocation upon termination. From (44) we obtain:

$$
|\hat{W}(\hat{a}) - \hat{W}(\hat{a}^*)| \leq \gamma N = \frac{\Delta^\text{min}}{4N}
$$

(45)

From lemma IV.9 and (45) we have $\hat{a} = \hat{a}^*$. 

The following corollary is a direct result of lemmas IV.7 and [4.10] is corollary IV.10.1.

Corollary IV.10.1. If $\varepsilon^* < \frac{\Delta^\text{min}}{8N}$ then the auction converges to $\hat{a}^*$ within $\frac{8N^3Q_M}{\Delta^\text{min}}(1 + \frac{1}{16N})$ iterations.

C. Proof of the Main Theorem

Proof: The number of epochs $J$ is $O(\ln T)$ since

$$
T \geq \sum_{j=1}^{J-1} 2^j = (2^J - 2).
$$

(46)

Hence, the accumulated regret of the exploration and auction phases of all the epochs of the algorithm is upper bounded by

$$
R_1 + R_2 \leq (T_1 + T_2)NQ_M \ln T
$$

(47)
since the regret at each slot is bounded by $N Q_M$. According to corollary [IV.10.1] the exploitation phase will only incur regret if the exploration or learning phases failed. According to corollary [IV.6.1] the probability of an exploration phase failure in epoch $j$ is upper bounded by

$$P\left(E^j\right) = O(s^{-j})$$

(48)

for some $s > 2$. Therefore, the accumulated regret of all the exploitation phases of all the epochs of the algorithm is upper bounded by

$$R_3 \leq N Q_M \sum_{j=1}^{J} (2/s)^j \leq C$$

(49)

which is bounded.

An interesting consequence of the main theorem is that the regret is mainly caused by the overhead of the learning phases, while the exploitation causes only bounded expected regret. In particular, when we implement a fixed size exploitation phase of length $T_3$ the regret will be bounded by

$$R \leq \frac{T_1 + T_2}{T_1 + T_2 + T_3} T + C$$

(50)

Setting $\alpha(T_1 + T_2) = T_3$ results in small linear regret of

$$R \leq \frac{1}{1 + \alpha} T + C$$

(51)

which the designer of the protocol can choose according to channel conditions and other practical considerations.

V. THE MEDIUM ACCESS CONTROL PROTOCOL

In practical scenarios using an exponentially increasing exploitation is unpractical since it will not allow the network to adapt to changing conditions quickly. Therefore we propose to replace the exponential learning phase with a fixed one, which consists of the larger portion of each epoch. We found that allocating $1-5\%$ of the resources to learning and coordination suffice to provide an accurate fast learning phase. We also recognized that initial learning is extremely important and propose to allocate 100ms to initialize the network from scratch where this phase consists of learning and coordination without exploitation. Once this phase ends, the network uses 5ms frames (which are approximately the typical coherence time) where 4$\mu$s will be used for channel exploration. 48$\mu$s will be used for auction and the rest for exploitation. The frame structure and the collision resolution state machines are as described in Figures [3][8] Table [III] presents the selected parameters for the network setup phase. Algorithm parameters such as for the auction are not changed during the steady state. Table [IV] defines the steady-state frame structure. As will be shown in simulations the following frame structure parameters provide a good starting point for channels with 5ms coherence time:

VI. SIMULATION RESULTS

In this section, we illustrate the performance of our proposed algorithm using computer simulations. The users of the network were located inside a disk with a radius of 100m. External interferers were located on a ring with an inner radius of 100m and an outer radius of 200m (see Figure [9]). The users in the inner disk implement our algorithm and their performance is measured, while the users in the outer ring have a static allocation and are treated as interference. To the south there is a strong interferer that is jamming half the frequencies in the southern part of the inner disk. The positions of some of the users and the strong interferer in an example scenario are shown in Figure [9].

Additionally, random $20\%$ of the remaining resource blocks were occupied by further external interference. We assume that the number of users is known in advance. However, this can be easily estimated distributedly by standard techniques.

The users had access to a bandwidth of 40 MHz around a central carrier frequency of 2GHz. The bandwidth is divided into 8 sub-bands of 5MHz each. The power spectral densities of the thermal noise and the signal are uniform over the entire bandwidth and equal to -174 dBm/Hz and -57 dBm/Hz, respectively. The set of transmission rates that the system allows is $\{1, 2, \ldots, 8\}$ bits per channel use. Each channel had 7 tap delays. The final tap delay has about one tenth of the power of the line-of-sight path. The channels have a Rayleigh distribution with a parameter $10^{-2}$. The shadowing effects are modelled by a log-normal distribution with a variance of 5 dBm. A summary of the channel parameters appears in table [I]. In all of the simulations the value of $\varepsilon$ at the end of the auction was $\varepsilon^* = \frac{1}{\lambda}$.

The first experiment demonstrate the regret bound with exponentially increasing exploitation phase. In figure [10] we validate our regret guarantee by simulating a fully-stationary network where the network is held fixed (including fading). 32 users are sharing 8 channels during 6 epochs with 1000 estimation blocks and 400 auction iterations in each epoch. Note that while the theoretical convergence guarantee for the auction is much larger, we observed that terminating the auction after 400-1000 iteration suffices to achieve very
Table I: Wireless channel model parameters.

| Channel Characteristic          | Value       |
|---------------------------------|-------------|
| carrier frequency               | 2GHz        |
| Total Bandwidth                 | 40MHz       |
| Sub-channel Bandwidth           | 5MHz        |
| path loss exponent $\alpha$     | 2           |
| Speed of light $c$              | $3 \cdot 10^6$ m/s |
| Distance between transceivers $n$ and $n'$ | $d_{n,n'}$ |
| Maximal tap delay between transceivers $n$ and $n'$ | $10^{25}d_{n,n'}/c$ |
| $\#$ uniformly distributed Tap Delays | 7          |
| Rayleigh Variance               | 0.01        |
| Shadow Log-Mean                 | 0           |
| Shadow Log-Variance             | 0.01        |
| Transmission power              | 1 mW        |
| Noise PSD                       | $-174$ dBm/Hz |
| Noise figure                    | 2 dB        |

Table II: MAC parameters of the regret plot

| MAC super-frame structure          | Value       |
|------------------------------------|-------------|
| Number of epochs                  | 6           |
| Expected time per epoch            | 4 ms        |
| Expected auction time per epoch    | 12 ms       |
| Number of possible discrete bids   | 256         |
| Initial $\epsilon$                | 1/32        |
| Final $\epsilon$                  | 1/32        |
| $\epsilon$- scaling factor        | 1           |

Table III: Setup phase parameter

| MAC super-frame structure          | Value       |
|------------------------------------|-------------|
| Exploration duration              | 80 ms       |
| Auction duration                  | 15 ms       |
| Auction iteration bound           | 500         |
| Expected number of collision-resolution blocks per frame | 0.36 |
| Initial epsilon initialization    | 1           |
| Final $\epsilon$                  | 1/32        |
| $\epsilon$- scaling factor        | 0.9808      |
| Number of possible discrete bids   | 4           |

We do not compare to multi-user reinforcement learning methods since these methods require passing information between users which is not allowed in our problem formulation.

The second experiment deals with a dynamic environment where channels vary. In this case a first stage of learning and acquisition is used, followed by fixed size epochs. The parameters of the setup stage appear in Table III and the parameters of the fixed-size epoch appear in Table IV.

Figures 11 and 12 show the performance of the algorithm with $100$ms of network setup followed by $100$ epochs of size $5$ms. In Figure 11 the network was assumed to be completely stationary. In Figure 12 the network was assumed to be dynamic such that the coherence time of the channel is $5$ms. The performance is averaged over 400 networks.

The performance (efficiency) is measured with respect to the optimal allocation computed by the Hungarian algorithm. We also present the performance of the greedy assignment (stable matching) and random allocations both using collision avoidance based on carrier sensing for computing the allocation (see e.g., [22]).

The proposed protocol achieves over $90\%$ efficiency in both cases. One can see a huge advantage of our algorithm compared to a random allocation which achieves in expectation less than $50\%$ efficiency. We also have a significant advantage over the greedy allocation. While the greedy allocation achieves in expectation $85\%$ efficiency in both the static and dynamic setting, our algorithm achieves $93\%$ efficiency in the dynamic setting and $95\%$ in the static setting. We have an advantage of $10\%$ over the greedy allocation in the static setting. Furthermore, we see that our $5\%$ outage rate is around $90\%$ for both the static and dynamic settings while the greedy algorithm achieves around $80\%$ efficiency relative to the optimal centralized solution.

Figures 13a and 13b shows a sample path of the efficiency of the proposed algorithm for a single dynamic network with $32$ users sharing $8$ channels over $100$ fixed-size epoch. Each epoch is $5$ms long. At most $1\%$ of the resources are dedicated to learning, so the maximal efficiency is around $99\%$. Indeed the efficiency converges to this value very quickly. Figure 13a shows the complete algorithm duration of network setup and $100$ steady state epochs. We see the rapid learning during the network setup and convergence to the maximum possible efficiency of $99\%$ during steady state. Figure 13b zooms on the steady state part of Figure 13a. Note that because of the dynamics of the channels, there are situations where efficiency is reduced until the protocol re-converges to the optimal allocation.

VII. CONCLUSION

We presented an algorithm for spectrum collaboration in congested ad-hoc networks that requires no direct communication between the users. Instead, the algorithm relies on local estimations of the channel state information and a distributed auction that is enabled by single-channel opportunistic carrier sensing. Because the number of users is greater than the number of channels, time-sharing is necessary. Therefore, spectrum collaboration must also include access scheduling.
Table IV: Steady state learning and exploitation structure

| MAC super-frame structure | Value       |
|---------------------------|-------------|
| Epoch duration            | 5ms         |
| Exploration time per epoch| 50µs        |
| Auction iterations per epoch | 4         |
| Expected auction phase duration | 200µs     |
| Exploitation phase        | 4750µs      |
| Number of possible discrete bids | 4           |
| Initial $\varepsilon$    | 1/32        |
| final $\varepsilon$      | 1/32        |
| $\varepsilon$-scaling factor | 1          |

Figure 11: Cumulative distribution function of the efficiency of the proposed algorithm after 100 fixed-sized epochs for static networks with 32 users and 8 sub-channels.

Our algorithm achieves this by learning the optimal OFDMA configuration with minimal overhead. Furthermore, with exponentially growing window our MAC protocol enjoys a $T$-optimal regret bound of $O(\ln T)$, and with a fixed duration exploitation phase, the protocol has an exponentially decaying probability of obtaining a sub-optimal action profile, while maintaining a fixed and low overhead. Simulations confirmed that our protocol significantly outperforms greedy/stable and random allocations over 5G channels. Further improvement can include the addition of RCS/CTS mechanism, to cope with hidden and exposed nodes. However, with the current signaling structure this is straightforward addition. Further optimization of the $\varepsilon$-scaling and the contention window size $\beta$ can speed up the convergence.

REFERENCES

[1] H. W. Kuhn, “The hungarian method for the assignment problem,” Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.
[2] J. Munkres, “Algorithms for the assignment and transportation problems,” Journal of the society for industrial and applied mathematics, vol. 5, no. 1, pp. 32–38, 1957.
[3] S. Chieochan and E. Hossain, “Adaptive radio resource allocation in OFDMA systems: A survey of the state-of-the-art approaches,” Wireless Communications and Mobile Computing, vol. 9, no. 4, pp. 513–527, 2009.
[4] E. Yaacoub and Z. Dawy, “A survey on uplink resource allocation in OFDMA wireless networks,” IEEE Communications Surveys & Tutorials, vol. 14, no. 2, pp. 322–337, 2011.
[5] F. Shams, G. Bacci, and M. Luise, “A survey on resource allocation techniques in OFDM(A) networks,” Computer Networks, vol. 65, pp. 129–150, 2014.
[6] H. Tibrewal, S. Patchala, M. K. Hanawal, and S. J. Darak, “Multiplayer multi-armed bandits for optimal assignment in heterogeneous networks,” arXiv preprint arXiv:1901.03868, 2019.
[7] A. Mehrabian, E. Boursier, E. Kaufmann, and V. Perchet, “A practical algorithm for multiplayer bandits when arm means vary among players,” Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 2020.
[8] P. Alatur, K. Y. Levy, and A. Krause, “Multi-player bandits: The adversarial case,” Journal of Machine Learning Research, vol. 21, 2020.
in nonstochastic bandits.” *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 613–650, 2019.

[14] C. Bowyer, D. Greene, T. Ward, M. Menendez, J. Shea, and T. Wong, “Reinforcement learning for mixed cooperative/competitive dynamic spectrum access,” in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*. IEEE, 2019, pp. 1–6.

[15] R. Batesley and R. Thompson, “Deep learning collaborative radios - team Zylinium phase 3;” *Defence Technical Information Center*, 2020.

[16] L. Perron, “Operations research and constraint programming at google,” in *International Conference on Principles and Practice of Constraint Programming*. Springer, 2011, pp. 2–2.

[17] R. Rom and M. Sidi, *Multiple access protocols: Performance and analysis*. Springer Science & Business Media, 2012.

[18] I. Bistritz and A. Leshem, “Distributed multi-player bandits-a game of thrones approach,” in *Advances in Neural Information Processing Systems*, 2018, pp. 7222–7232.

[19] ———, “Game of thrones: Fully distributed learning for multi-player bandits,” *Mathematics of operations*, 2020.

[20] W. Wang, A. Leshem, D. Niyato, and Z. Han, “Decentralized learning for channel allocation in iot networks over unlicensed bandwidth as a contextual multi-player multi-armed bandit game,” *arXiv preprint arXiv:2003.13314*, 2020.

[21] I. Bistritz, T. Baharav, A. Leshem, and N. Bambos, “My fair bandit: Distributed learning of max-min fairness with multi-player bandits,” *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[22] A. Leshem, E. Zehavi, and Y. Yaffe, “Multichannel opportunistic carrier sensing for stable channel access control in cognitive radio systems,” *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 1, pp. 82–95, 2011.

[23] S. J. Darak and M. K. Hanawal, “Multi-player multi-armed bandits for stable allocation in heterogeneous ad-hoc networks,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2350–2363, 2019.

[24] O. Avner and S. Mannor, “Multi-user communication networks: A coordinated multi-player bandit approach,” *IEEE/ACM Transactions on Networking*, vol. 27, no. 6, pp. 2192–2207, 2019.

[25] N. Nayyar, D. Kalathil, and R. Jain, “On regret-optimal learning in decentralized multi-player multiarmed bandits,” *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 597–606, 2016.

[26] O. Naparstek and A. Leshem, “Fully distributed optimal channel assignment for open spectrum access,” *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 283–294, 2013.

[27] S. Zafaruddin, I. Bistritz, A. Leshem, and D. Niyato, “Distributed learning for channel allocation over a shared spectrum,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2337–2349, 2019.

[28] O. Naparstek, S. M. Zafaruddin, A. Leshem, and E. A. Jorswieck, “One for all and all for one: Distributed learning of fair allocations with multi-player bandits,” *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 4, pp. 1152–1166, 2016.

[29] T. Gafni and K. Cohen, “Distributed learning over markovian fading channels for stable spectrum access,” *arXiv preprint arXiv:2101.11922*, 2021.

[30] O. Naparstek, T. Z. Baharav, A. Leshem, and N. Bambos, “One for all and all for one: Distributed learning of fair allocations with multi-player bandits,” *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 584–598, 2021.

[31] H. Liu, K. Liu, and Q. Zhao, “Learning in a changing world: Restless bandit problems and optimality of whittle index for dynamic multi-channel access,” *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2012.

[32] K. Cohen, Q. Zhao, and A. Scaglione, “Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access,” in *2014 48th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2014, pp. 1575–1578.

[33] K. Liu and Q. Zhao, “Indexability of restless bandit problems and optimality of whittle index for dynamic multi-channel access,” *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[34] M. Yemini, A. Leshem, and A. Somekh-Baruch, “The restless hidden markov bandit with linear rewards and side information,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 1108–1123, 2021.

[35] O. Naparstek and K. Cohen, “Deep multi-user reinforcement learning for distributed dynamic spectrum access,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 310–323, 2018.

[36] P. M. Pawar and A. Leshem, “Distributed deep reinforcement learning for collaborative spectrum sharing,” *arXiv preprint arXiv:2104.02059*, 2021.

[37] Y. Yu, T. Wang, and S. C. Liew, “Deep-reinforcement learning multiple access for heterogeneous wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1277–1290, 2019.

[38] H. Q. Nguyen, T. Q. Dong, D. T. Ngo, and T. A. Nguyen, “Deep q-learning with multiband sensing for dynamic spectrum access,” in *2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2018, pp. 1–5.

[39] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li, “Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach,” *IEEE access*, vol. 6, pp. 25 463–25 473, 2018.

[40] Y. Yu, T. Wang, and S. C. Liew, “Deep-reinforcement learning multiple access for heterogeneous wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1277–1290, 2019.

[41] E. Koutsoupias and C. Papadimitriou, “Worst-case equilibria,” in *Annual Symposium on Theoretical Aspects of Computer Science*. Springer, 1999, pp. 404–413.

[42] J. R. Marden, H. P. Young, and L. Y. Pao, “Achieving pareto optimality through distributed learning,” *SIAM Journal on Control and Optimization*, vol. 52, no. 5, pp. 2753–2770, 2014.

[43] Q. Zhao and L. Tong, “Opportunistic carrier sensing for energy-efficient information retrieval in sensor networks,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2005, no. 2, pp. 1–11, 2005.

[44] E. A. Jorswieck, “Stable matchings for resource allocation in wireless networks,” in *2011 17th International Conference on Digital Signal Processing (DSP).* IEEE, 2011, pp. 1–8.

[45] B. Holfeld and E. Jorswieck, “On stable many-to-many matching for distributed medium access with reuse of spectral resources,” in *WSA 2016; 20th International ITG Workshop on Smart Antennas*. VDE, 2016, pp. 1–8.

[46] T. Höller, P. Schulz, E. A. Jorswieck, M. Simsek, and G. P. Fettweis, “Stable matching for wireless urllc in multi-cellular, multi-user systems,” *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5228–5241, 2020.

[47] Z. Han, Y. Gu, and W. Saad, “Stable marriage model with cheating for d2d communications,” in *Matching Theory for Wireless Networks*. Springer, 2017, pp. 17–37.

[48] Y. Xiao, D. Niyato, K.-C. Chen, and Z. Han, “Enhance device-to-device communication with social awareness: A belief-based stable marriage game framework,” *IEEE Wireless Communications*, vol. 23, no. 4, pp. 36–44, 2016.

[49] T. P. Bertsekas, “A distributed algorithm for the assignment problem,” *Lab. for Information and Decision Systems Working Paper, MIT*, 1979.

[50] D. J. Aldous, “The 3 (2) limit in the random assignment problem,” *Random Structures & Algorithms*, vol. 18, no. 4, pp. 381–418, 2001.

[51] N. Alon and J. H. Spencer, *The probabilistic method*. John Wiley & Sons, 2004.

---

**Appendix**

### A. Detailed Proof of Lemma IV.4

**Proof:** To prove the lemma it is sufficient to prove that

\[ \xi \leq \xi^* \Rightarrow -E^j. \]

Let \( a’ \) be a sub-optimal action profile. By our assumption on \( Q_{n,k} \) we have:

\[ W(a^*) - W(a') \geq \Delta_{min}. \]  

(52)

From (25) and (26) we obtain:

\[ W^j(a^*) \geq W(a') - (\xi + \Delta_{max})N \]

(53)

\[ W^j(a') \leq W(a') + \xi N \]

(54)

Substitute (53) and (54) into (52):

\[ W^j(a^*) - W^j(a') \geq \Delta_{min} - 2N - \frac{\Delta_{min}}{4} > 0 \]

(55)

Where (a) holds for \( \xi < \xi_{max} \).
**B. Detailed Proof of Lemma IV.5**

Proof:

Let $H$ be the number of Heads in $n$ flips of a coin with probability of $p$. The standard form of Hoeffding’s inequality for Bernoulli random variables with probability of success $p$ and number of trials $n$ is:

$$P(H \leq (p + \epsilon)n) \leq \exp(-2n\epsilon^2) \quad (56)$$

Here the number of trials is $10jN^3\Delta^{-2}$ and the probability of success is

$$P_{ss} \triangleq \frac{1}{N} \left(1 - \frac{1}{N}\right)^{N-1} \quad (57)$$

We re-write Hoeffding’s inequality with these parameters:

$$P(H \leq (P_{ss} + \epsilon)10jN^3\Delta^{-2}) \leq \exp(-20jN^3\Delta^{-2}\epsilon^2) \quad (58)$$

We set:

$$(P_{ss} + \epsilon)10jN^3\Delta^{-2} = V_{\text{min}} \quad (59)$$

Therefore:

$$\epsilon = \frac{V_{\text{min}}}{10jN^3\Delta^{-2}} - P_{ss} \quad (60)$$

Hoeffding’s inequality becomes

$$P(V \leq V_{\text{min}}) \leq \exp(-20jN^3\Delta^{-2}(\frac{V_{\text{min}}}{10jN^3\Delta^{-2}} - P_{ss})^2) \quad (61)$$

Set $P_{ss}$ and

$$V_{\text{min}} \triangleq \frac{5jN^2}{2\Delta^2} \quad (62)$$

into the inequality to obtain

$$P(V \leq V_{\text{min}}) \leq \exp\left(-20jN^3\Delta^{-2}\left(\frac{1}{4} - \left(1 - \frac{1}{N}\right)^{N-1}\right)^2N^{-2}\right) \quad (63)$$

We have

$$\left(1 - \frac{1}{N}\right)^{N-1} = 0.25 \geq e^{-1} - 0.25 \geq 1/9 \quad (64)$$

Set this into Hoeffding:

$$P(V \leq V_{\text{min}}) \leq \exp(-20jN^3\Delta^{-2}/81N^2) \quad (65)$$

After the union bound on the users and channels we have

$$P(E^j_V) < NK \exp(-20jN\Delta^{-2}/81) \quad (66)$$

If the following condition holds

$$20jN\Delta^{-2}/81 > \ln(2) \quad (67)$$

then we obtain that the error probability decreases like

$$P(E^j_V) = o(2^{-j}) \quad (68)$$

Simple algebra show that when $\Delta_{\text{min}} = 1$, (67) holds as long as $N \geq 2$ and $Q_M \geq 2$.

**C. Detailed Proof of Lemma IV.6**

Proof: Let $X_1, \ldots, X_n$ be bounded random variables such that $X_i$ is bounded between $a_i$ and $b_i$. Let $X \triangleq \frac{1}{n}(X_1 + \ldots + X_n)$. According to Hoeffding’s inequality for bounded random variables

$$P(|X - E(X)| > t) < 2\exp\left(-\frac{n^2t^2}{\sum_i(b_i - a_i)^2}\right) \quad (69)$$

In our case $X_1, \ldots, X_n$ are i.i.d. samples of a channel, the upper bound is $\forall i, b_i = Q_M$, the lower bound is $\forall i, a_i = 0$, the allowed error is $t = \xi_{\text{max}}$, the number of samples is $n = V^j_{\text{min}}$. Setting all these definitions:

$$P\left(E^j \mid -E^j_V\right) \leq 2NK \exp\left(-2(\frac{\xi_{\text{max}}(V^j_{\text{min}})^2}{Q^2_M V^j_{\text{min}}}\right) \quad (70)$$

Because: $\frac{(V^j_{\text{min}})^2}{V^j_{\text{min}}} = V^j$ we have:

$$P\left(E^j \mid -E^j_V\right) \leq 2NK \exp\left(-2\frac{\xi_{\text{max}}V^j_{\text{min}}}{Q^2_M}\right) \quad (71)$$

Substituting $\xi_{\text{max}} = \frac{3\Delta_{\text{min}}}{8N}$ we obtain:

$$P\left(E^j \mid -E^j_V\right) \leq 2NK \exp\left(-2\left(\frac{3\Delta_{\text{min}}}{8N}\right)^2 V^j_{\text{min}}/Q^2_M\right) \quad (72)$$

Substituting $V^j_{\text{min}} = \frac{5jN^2}{2\Delta^2}$ and simple algebra we obtain:

$$P\left(E^j \mid -E^j_V\right) \leq 2NK \exp\left(-\frac{45jN^3\Delta^{-2}/64N^3}{64}\right) \quad (73)$$

$$= 2NK \exp\left(-\frac{45j}{64}\right)$$

Because $45/64 \geq \ln(2)$ we obtain that the error probability decreases like

$$P(E^j \mid -E^j_V) = o(2^{-j}) \quad (74)$$

**D. Detailed Proof of Lemma IV.9**

Proof: For this proof we use the following abbreviations:

$W = W(a), W' = W(a'), \hat{W} = \hat{W}(a), \hat{W}' = \hat{W}(a')$

The absolute difference between the estimated welfares of $a$ and $a'$ is bounded from below by

$$|\hat{W} - \hat{W}'| = |W - W' + \hat{W} - W + W' - \hat{W}'| \quad (a)$$

$$\geq |W - W'| - |\hat{W} - W| - |W' - \hat{W}'| \quad (75)$$

$$\geq \Delta_{\text{min}} - \frac{3\Delta_{\text{min}}}{4} \geq \frac{\Delta_{\text{min}}}{4}$$

where (a) follows from the reverse triangle inequality and (b) from the definition of $Q_{n,k}$ and the success of the $j$-th exploration phase that results in $\xi < \frac{3\Delta_{\text{min}}}{8N}$. 

