Multiagent Autonomous Learning for Distributed Channel Allocation in Wireless Networks

Syed Mohammad Zafaruddin*†, Ilai Bistritz*, Amir Leshem*, and Dusit Niyato‡
* Faculty of Enginerring, Bar-Ilan University, Ramat Gan 52900, Israel, leshema@eng.biu.ac.il
† Deptt. of Electrical and Electronics Engineering, BITS Pilani, Rajasthan, India, syed.zafaruddin@pilani.bits-pilani.ac.in
‡School of Computer Science and Engineering (SCSE), Nanyang Technological University, Singapore, dniyato@ntu.edu.sg

***Abstract*—In distributed networks such as ad-hoc and device-to-device (D2D) networks, no base station exists and conveying global channel state information (CSI) between users is costly or simply impractical. When the CSI is time-varying and unknown to the users, the users face the challenge of both learning the channel statistics online and converging to good channel allocation. This introduces a multi-armed bandit (MAB) scenario with multiple decision makers. If two or more users choose the same channel, a collision occurs and they all receive zero reward. We propose a distributed channel allocation algorithm in which each user converges to the optimal allocation while achieving an order optimal regret of $\mathcal{O}\left(\log T\right)$, where T denotes the length of time horizon. The algorithm is based on a carrier sensing multiple access (CSMA) implementation of the distributed auction algorithm. It does not require any exchange of information between users. Users need only to observe a single channel at a time and sense if there is a transmission on that channel, without decoding the transmissions or identifying the transmitting users. We compare the performance of the proposed algorithm with the state-of-the-art scheme using simulations of realistic long term evolution (LTE) channels.**

***Index Terms*—Distributed channel allocation, multiplayer multi-armed bandit, online learning, dynamic spectrum access, resource management, wireless networks.**

I. Introduction

Channel allocation in wireless communication is one of the fundamental management tasks [1], [2]. In the traditional centralized systems which have global view of the whole network, the optimal solution of a desired performance metric can be obtained using the Hungarian method. However, emerging wireless networking paradigms such as cognitive radio networks, ad-hoc networks, and device-to-device (D2D) communications are inherently distributed. A complete information about the network state for these networks is typically not available online, which makes the computation of optimal policies intractable. Hence, it is desirable to develop a distributed learning algorithm for dynamic spectrum access that can effectively adapt to general complex real-world settings in dense and heterogeneous wireless environments.

The center of the channel allocation task is the combinatorial optimization assignment problem. The literature on distributed channel allocation without learning, where the channel state information (CSI) is assumed to be known, is vast and we can only cover part of it here. Recently there has been growing interest in distributed spectrum optimization for frequency selective channels, where the assignment problem arises. However, most of the work done in this field relies on explicit exchange of CSI. Several suboptimal approaches that do not require information sharing have been suggested [3]–[6].

The auction algorithm [7] has been extensively used to solve a variety of assignment problems for channel allocation [8]–[11]. In [8] the auction algorithm was used to solve the channel assignment problem for the uplink, using the base station as the auctioneer. In [9] a distributed auction algorithm with shared memory was used for switch scheduling. In [10] it was shown that a modification of the auction algorithm is equivalent to max product belief propagation. However, all these modified auction algorithms require a base station or shared memory, which prevents them from being fully distributed. In [11] a fully distributed version of the auction algorithm was suggested that exploits carrier sense multiple access (CSMA) in order to avoid the need for an auctioneer. In addition, all these algorithms, including [11] that is being used here, assume that the CSI is known to the users. Our algorithm generalizes the distributed CSMA auction algorithm [11] to an online learning framework.

If the resource (channel) values are not known in advance by the users, they have to learn these values online. Learning the CSI in real-time comes at the expense of using the best known channels so far. This introduces the well-known trade off between exploration and exploitation that is captured by the multi-agent multi-armed bandit (MAB) problem. Developing MAB-based methods for solving dynamic spectrum allocation problems is an interesting research direction, motivated by recent developments of MAB in various other fields, and many works have been done in this direction recently. A couple of these works [12]–[14] considered a cognitive radio scenario where a set of channels can be either free or occupied by a primary user that interferes all secondary users. A generalized scenario was considered in [15]–[17], where the channel qualities are not binary, but still all users have the same vector of channel qualities. Recently, the case of a full channel allocation scenario where different users have different channel qualities (a matrix of channel qualities) was considered in [18]. Later the channel allocation was improved in [19] by the same authors, to have an order optimal sum-regret of $\mathcal{O}\left(\log T\right)$, where T denotes the length of time horizon. In [19], the auction algorithm [7] was used as a basis for a distributed algorithm that achieves an expected sum regret of $\mathcal{O}(\log T)$. However, since it relies on [7], this algorithm requires communication between users in order to exchange the bids and determine the winning player in each auction.

Recently, it has been shown in [20] (which improved [21]) that achieving a sum-regret of near-$\mathcal{O}(\log T)$ is possible even without communication between users and with a matrix of expected rewards. The algorithm in [20] is general but has a slow convergence rate in T that makes it unsuited for realistic communication scenarios. In this paper, we adopt a more practical and communication-oriented approach and achieve an order optimal sum-regret of $\mathcal{O}(\log T)$. The proposed algorithm does not require any communication between users, and each user only needs to sense a single channel at a time (instead of simultaneously all of them as in [19]). It is made possible by adding assumptions that are always valid from a practical perspective: the expected rewards are integer multiplications of a common resolution, and a user can choose not to transmit on any channel and instead can sense a single channel of its choice. We show that the proposed algorithm is much easier and less costly to implement than that of [19] and has a much better convergence time than that of [20].

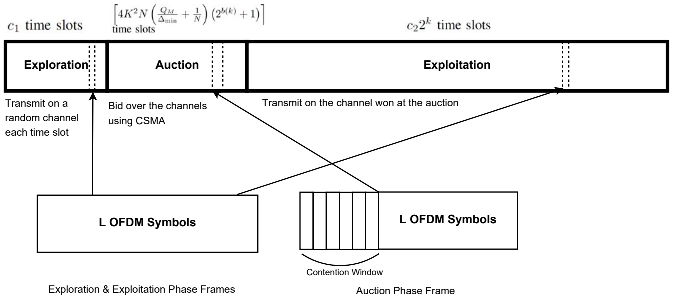
II. System Model

We consider an ad hoc network with a set of transmitter-receiver pairs (links) $\mathcal{N} = \{1, \ldots, N\}$ and a set of channels $\mathcal{K} = \{1, \ldots, K\}$, where $K \geq N$. Each channel consists of several orthogonal frequency-domain multiple access (OFDMA) subcarriers, and each link uses a single channel. In the case of more users than channels ($N > K$), a combined OFDMA-TDMA (time division multiple access) can be used instead in order to have enough resources for all users. The links are located in a geographical proximity in an area that typically includes other coexisting networks nearby. This is relevant, for example, for WiFi networks and Internet of Things (IoT) networks. As a result, each receiver can experience alien interference from the transmission of other users. Time is slotted and indexed by t, such that in each time slot, L OFDM symbols are transmitted. We assume a fast-fading scenario such that the number of OFDM symbols per time slot L is designed to match the coherence time of the channel. The links are active for a total of T time slots, where T is unknown in advance by the links. The chosen channel of link n at time t is denoted by $a_n(t)$. Naturally, links can choose not to transmit at all at a given time slot, which is denoted $a_n(t) = 0$. Non-transmitting links can still sense transmissions on a single chosen channel.

Since the channel statistics and the interference pattern are initially unknown, each link needs to learn them online as fast as possible in order to deduce which quality of service (QoS) that it can support. The supported QoS set is $\mathcal{Q} \triangleq \{Q_1, \cdots, Q_M\}$ where for each i, $Q_i = l_i \Delta_{\min}$, where $\Delta_{\min}$ is the resolution for the supported QoS for a non-negative integer l_i and $Q_1 < \cdots < Q_M$. The QoS experienced by link n using channel i is denoted by $Q_{n,i}$. In general, different links have a subset of different possible QoS values from $\mathcal{Q}$ due to different capabilities, e.g., number of transmitting and receiving antennas. Being part of the standard of the protocol, we assume that the parameters $\Delta_{\min}$ and $\Delta_{\max} = Q_M - Q_1$ are known to all devices.

In each time slot t, each link measures the instantaneous QoS $q_{n,i}(t)$ by using a finer resolution than that of $\mathcal{Q}$, in order for the estimation of the average to be accurate. We model $q_{n,i}(t)$ as i.i.d. sequence in time, independent for different n or i. The distribution of $q_{n,i}(t)$ is bounded since $Q_1 \leq q_{n,i}(t) \leq Q_M$, and can be either discrete or continuous due to arbitrarily fine measurements. We define the no-collision indicator of channel i at time t by $\eta_i = 0$ when $|\mathcal{N}_i(t)| > 1$ and $\eta_i = 1$ otherwise, where $\mathcal{N}_i$ denotes the set of links that are transmitting on channel i at time t. The instantaneous reward of link n at time t from transmitting on channel a_n is

$$r_{n,a_n}(t) = q_{n,a_n}(t)\, \eta_{a_n}(t) \,. \tag{1}$$



c_1 time slots
$\left\lceil 4K^2N\left(\frac{Q_M}{\Delta_{\min}}+\frac{1}{N}\right)\left(2^{b(k)}+1\right)\right\rceil$ time slots
$c_2 2^k$ time slots
Exploration
Auction
Exploitation
Transmit on a random channel each time slot
Bid over the channels using CSMA
Transmit on the channel won at the auction
L OFDM Symbols
L OFDM Symbols
Contention Window
Exploration & Exploitation Phase Frames
Auction Phase Frame

Fig. 1. The k-th packet structure of the proposed algorithm.

For theoretical evaluation of our algorithm, we use the regret which is defined as

$$R = \sum_{t=1}^{T}\sum_{n=1}^{N} Q_n^* - \sum_{t=1}^{T}\sum_{n=1}^{N} q_{n,a_n(t)}(t)\, \eta_{a_n(t)}(t) \,. \tag{2}$$

The value Q_n^* is the expectation of the QoS of the channel that link n is assigned to:

$$a^* = \arg\max_{a_1,\ldots,a_N} \sum_{n=1}^{N} Q_{n,a_n} \,. \tag{3}$$

The expected total regret $\bar{R} \triangleq E\{R\}$ is the average of (2) over the randomness of the rewards $\{r_{n,i}(t)\}_t$ that dictate the random channel choices $\{a_n(t)\}$.

III. Proposed Protocol

We design a novel multi-access channel (MAC) protocol based on distributed auction algorithm where each link runs distributively in order to maximize the accumulated sum of QoS. It is noted that the algorithm in [11] exploits the CSMA mechanism to bypass the need for an auctioneer and by doing that, implements the auction algorithm distributively. For this purpose, links compute a continuous back-off time that is decreasing with their bid. The highest bidder for a particular channel is simply the first link which accesses this channel. However, in contrast to [11], we assume all links can sense the channel that they choose, and all links will agree on which link is the highest bidder for their channel. *Note that we do not analyze selfish links, but consider devices that are programmed to run our designed MAC protocol in a cooperative manner. This is the way that most MAC protocols operate.*

We divide the T time slots into packets with a dynamic length, one starting immediately after the other. Each packet is further divided into three phases, as shown in Fig. 1. In the k-th packet:

1) **Exploration Phase** — this phase has a length of c_1 time slots in each packet, and is used for estimating the expected reward in each channel. The estimated values are artificially dithered in order to avoid ties in the subsequent auction phase. Collisions can be excluded from each link's estimation since they result in zero reward.
2) **Auction Phase** — this phase has a length of $\left\lceil 4K^2N\left(\frac{Q_M}{\Delta_{\min}}+\frac{1}{N}\right)\left(2^{b(k)}+1\right)\right\rceil$ time slots in the k-th packet, which is the convergence time of the distributed auction algorithm, as dictated by Lemma 3. In this phase, the links run the distributed auction on the estimated

```
Algorithm 1 Distributed Learning for Channel Allocation
Initialization Choose $\varepsilon < \frac{\Delta_{\min}}{4K}$. Set $V_{n,i}(0) = 0$ and $s_{n,i}(0) = 0$ for all $i$ and $b(0) = 8$.
  1) Dither Values — Generate $u_{n,i}$ for each $i$, independently and uniformly distributed over $\left[-\frac{\Delta_{\min}}{8N}, \frac{\Delta_{\min}}{8N}\right]$.
For $t = 1, \ldots, T$ (which determines $k$) do
A. Exploration Phase — For the next $c_1$ time slots
  1) Choose a channel $i \in [1, .., K]$ uniformly at random.
  2) Receive the reward $r_{n,i}(t)$. Update $V_{n,i}(t) = V_{n,i}(t-1) + \eta_i(t)$ and $s_{n,i}(t) = s_{n,i}(t-1) + r_{n,i}(t)$, where $\eta_i$ and $r_{n,i}$ are defined in Section II.
  3) Create a dithered estimation of $Q_{n,i}$ by computing $Q^k_{n,i} = \frac{s_i(t)}{o_i} + u_{n,i}$ for $i = 1, \ldots, K$.
B. Auction Phase — set state unassigned and $B_{n,i} = 0, \forall i$.
For the next $\left\lceil 4K^2N\left(\frac{Q_M}{\Delta_{\min}} + \frac{1}{N}\right)\left(2^{b(k)}+1\right)\right\rceil$ time slots
Each auction iteration do
  1) If unassigned then
    a) Calculate its own maximum profit:
         $\gamma_n = \max_i \left(Q^k_{n,i} - B_{n,i}\right)$    (4)
    b) Calculate its own second maximum profit:
         $\tilde{i}_n = \arg\max_k \left(Q^k_{n,i} - B_{n,i}\right)$    (5)
         $w_n = \max_{i \neq \tilde{i}_n} \left(Q^k_{n,i} - B_{n,i}\right)$    (6)
    c) Update the bid for its best channel $\tilde{i}_n$:
         $B_{n,\tilde{i}_n} = B_{n,\tilde{i}_n} + \gamma_n - w_n + \varepsilon$    (7)
  2) During the next $2^{b(k)}$ time slots — Sense the channel $\tilde{i}_n$ after a back-off time of
         $\tau_n = f_{b(k)}\left(B_{n,\tilde{i}_n}\right)$    (8)
     time slots, where $f_{b(k)}$ is a quantization of some decreasing function $f$ (e.g., $f(x) = 2^{b(k)} - x$) using $b(k)$ bits, such that $0 \le \tau_n \le 2^{b(k)}$.
    a) If the channel is not busy set state to assigned and to unassigned otherwise.
  3) Collision Resolution — In the $\tau_{\max} = 2^{b(k)} + 1$ time slot
    a) Transmit over channel 1 if it is assigned a channel with a collision.
    b) If links sense a transmission on channel 1, then they update $b(k+1) = b(k) + 1$.
End
C. Exploitation Phase — for the next $c_2 2^k$ time slots
  a) Transmit over the channel assigned at the end of the auction phase.
End
```

expected rewards using $b(k)$ bits for the quantized back-off time. The function $b(k)$ converges to a constant that is independent of k.

3) **Exploitation Phase** — this phase has a length of $c_2 2^k$ time slots for some constant c_2. During this phase, the links transmit on the channel that they are allocated in the auction phase.

The description of algorithm is given in Algorithm 1.

The key advantage of our algorithm is that it only requires from each receiver to sense if there are transmissions on a single channel, which is a basic requirement. We assume that all links are at a sensing distance from each other (a fully-connected network). As is common in CSMA systems, this assumption can be relaxed using request to send or clear to send (RTS/CTS) protocols where the RTS/CTS are much shorter and have priority in access. However, for simplicity of exposition, we ignore this aspect. However, as opposed to [19], the links do not know which transmission belongs to which link. This is the scenario in practice with wireless links located in close enough proximity. In our protocol, links do not need to distinguish between the transmission of other links, which may require decoding an ID for each link. Moreover, it can be extremely computationally demanding in practice to separate colliding transmissions and discern the IDs involved. Sensing a single channel at a time instead of all the K channels is another major advantage of our algorithm over [19].

The fact that the exploitation phase requires an exponential number of time slots does not mean that it takes longer time—it means that the lengths of the exploration and auction phases are relatively much shorter. Note that T is finite and can be set by the designer. Therefore, even the last (longest) exploitation phase can still consist of just a couple of thousands of OFDM symbols, which amounts to only a few milliseconds. From a practical point of view, this is the desirable packet structure since the actual transmission takes the vast majority of the OFDM symbols while the equivalents of the synchronization header do not cause a significant overhead. The overhead caused by the exploration and auction phases is naturally measured by the sum of regrets as in (2). We also note that the computational complexity of running Algorithm 1 for each device is $\mathcal{O}(K)$, since maximization over a K-sized vectors is required.

IV. Regret Performance Analysis

In this section, we analyze the performance of the proposed protocol in different phases and provide our main result: The expected sum of regret is an order optimal regret of $\mathcal{O}(\log T)$.

The following lemma characterizes the required estimation accuracy of the exploration phase, taking into account the dither noise.

Lemma 1 (Accuracy of Exploration Phase)**.** *Denote the dithered estimations of the expected QoS values in packet k by $\left\{Q^k_{n,i}\right\}$. Assume that $\left|Q^k_{n,i} - Q_{n,i} - u_{n,i}\right| \le \Delta$ for each link n and channel i for some positive Δ. If $\Delta < \frac{3\Delta_{\min}}{8N}$, then*

$$\arg\max_{a_1,\ldots,a_N} \sum_{n=1}^{N} Q_{n,a(n)} = \arg\max_{a_1,\ldots,a_N} \sum_{n=1}^{N} Q^k_{n,a(n)}. \qquad (9)$$

Proof: The proof follows from the fact that if $Q^k_{n,i}$ and $Q_{n,i}$ are close enough for every i and n, then the optimal assignment on $\left\{Q^k_{n,i}\right\}$ and $\left\{Q_{n,i}\right\}$ must be identical. For details see the extended version of the paper [22]. ■

The following lemma provides an upper bound for the probability that the estimation for packet k failed. The fact that this error probability exponentially vanishes with k, allows us to limit the number of exploration time slots to c_1, keeping the overhead caused by the exploration phase negligible.

Lemma 2 (Exploration Error Probability)**.** *Denote the dithered estimations of the expected QoS values in packet k by $\left\{Q^k_{n,i}\right\}$. If the length of the exploration phase satisfies*

$c_1 \geq K \max\left\{\frac{81}{2}K, \frac{128}{9}\left(\frac{\Delta_{\max}}{\Delta_{\min}}\right)^2 N^2\right\}$*, then after the k-th packet, we have*

$$P_{e,k} \triangleq \Pr\left(\max_{n,i}\left|Q_{n,i}^k - Q_{n,i}\right| > \frac{3\Delta_{\min}}{8N}\right) \leq 3NKe^{-k}. \tag{10}$$

Proof: The proof uses Hoeffding's bound on both $\left|Q_{n,i}^k - Q_{n,i}\right|$ and the number of samples of $Q_{n,i}$ without collision. For details see the extended version of the paper [22]. ■

Lemma 3 (Auction Phase)**.** *Assume that $b(k') = b_f$ for all $k' > k$. If the k-th exploration phase succeeded, then the k-th auction phase converges to an allocation $a_1, \ldots, a_N$ such that $\left|\sum_{n=1}^N Q_{n,a_n}^k - \max_{a_1,\ldots,a_N} \sum_{n=1}^N Q_{n,a_n}^k\right| \leq \varepsilon$ with less than $\frac{KN}{\varepsilon_k}\left(Q_M + \frac{\Delta_{\min}}{8N}\right)2^{b(k)}$ time slots with the probability 1. If $\varepsilon < \frac{3\Delta_{\min}}{4K}$, and then the auction phase converges to $\arg\max_a \sum_{n=1}^N Q_{n,a_n}$.*

Proof: The proof follows from the convergence and performance guarantees proven in [23] together with Lemma 2. For details see the extended version of the paper [22]. ■

Finally, we present our main result in the following theorem.

Theorem 1 (Main Result)**.** *Assume that the instantaneous QoS $\{q_{n,i}(t)\}_t$ are independent in n and i.i.d in time t, with expectations $Q_{n,i} \in \{Q_1, \ldots, Q_M\}$ such that $Q_i = l_i \Delta_{\min}$ for a non-negative integer l_i and a positive $\Delta_{\min}$, and $Q_1 < \ldots < Q_M$. Denote $\Delta_{\max} = Q_M - Q_1$. Let each link run Algorithm 1 with $\varepsilon < \frac{\Delta_{\min}}{4K}$ and an exploration phase length of*

$$c_1 \geq K \max\left\{\frac{81}{2}K, \frac{128}{9}\left(\frac{\Delta_{\max}}{\Delta_{\min}}\right)^2 N^2\right\}. \tag{11}$$

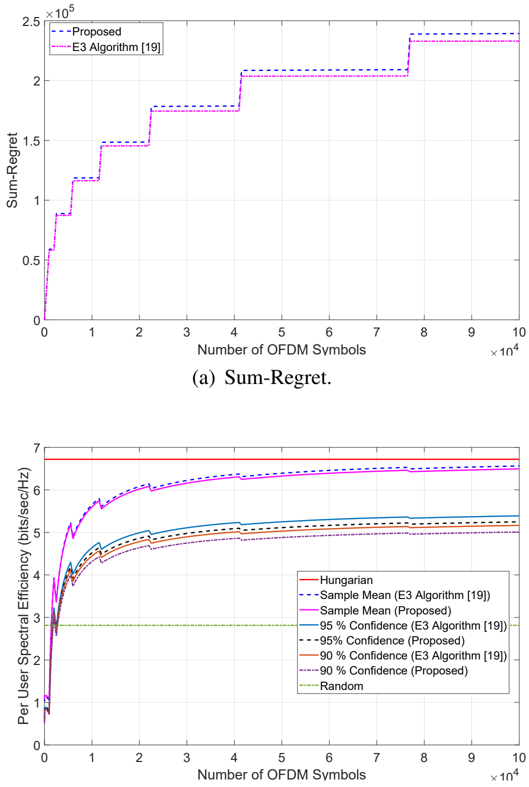
Then, the expected sum of regrets is $\bar{R} \sim \mathcal{O}(\log T)$.

Proof: Lemma 3 shows that if the exploration phase succeeds and enough bits are used for the CSMA back-off quantization, then the exploitation phase contributes no regret to the sum of regret. Moreover, Lemma 2 upper bounds from above the error probability of the exploration phase. The bound implies that it decreases exponentially with k. The proof follows by bounding from above the expected regret using these two facts. The proof is presented in the extended version of the paper [22]. ■

V. Simulation Results

In this section, we demonstrate the performance of proposed algorithm using computer simulations. We compare the proposed algorithm with the centralized Hungarian method [24], random channel selection and the E3 algorithm in [19]. The Hungarian method requires some central entity to know the CSI of all users. Requiring much less information, the E3 algorithm assumes that each user can decode the channel each of the other users choose. Our algorithm requires even much less information - each user only needs to sense whether there is a transmission on a given channel.

We consider an ad-hoc network of N links that are uniformly distributed on disk with a radius of 500 m. We considered the extended pedestrian A model (EPA) of the LTE standard with 9 random taps. The path loss is computed using path loss exponent of $\alpha = 4$. The central carrier frequency is 2 GHz with a per-user transmission bandwidth of 200 KHz. The channel bandwidth is divided into N sub-channels and we used $N = K = 10$. The transmit power is fixed at 12dBm for each user. The users were assumed to be moving at a speed of 3 km/h. We used a transmission duration of $T = 10^5$ time slots, with a single OFDM symbol per time slot ($L = 1$). Our transmission packet has the exploration phase of 800 OFDM symbols and the auction phase of 500 OFDM symbols. Each experiment consists of averaging 1000 independent realizations.



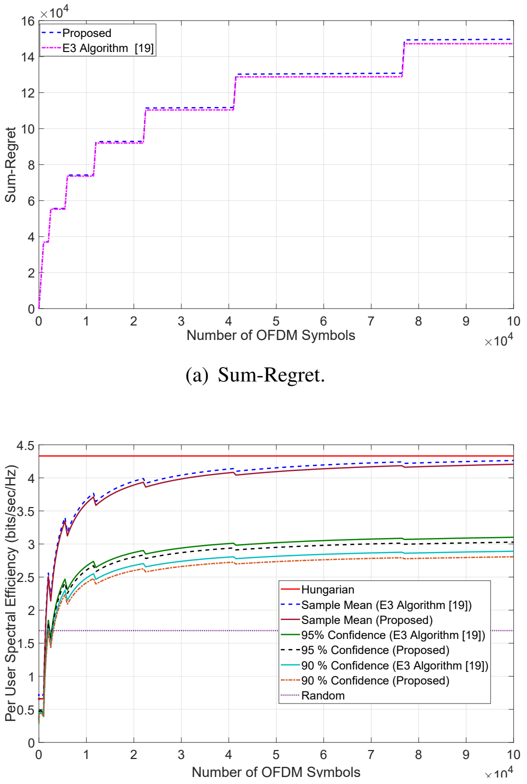
(a) Sum-Regret.

(b) Spectral Efficiency.

Fig. 2. Performance evaluation over LTE fading channel. Simulation parameters are: $N = K = 10$, explore length= 800 OFDM symbols, and auction length = 500 OFDM symbols.

In Fig. 2a, the sum-regret of our algorithm is compared with that of the E3 algorithm [19]. It is evident that the performance of both algorithms is essentially identical, despite the fact that our algorithm uses no communication between users as the E3 algorithm [19] does. Both algorithms have an expected sum-regret that increases similar to $\log T$ and both converge to the optimal allocation already at the first packets. In Fig. 2b, we present the spectral efficiency performance of both algorithms together with the confidence intervals of 90% and 95% outage (worst realizations)), where again all performances are very similar between our algorithm and the E3 algorithm [19]. It also shows that the proposed algorithm approaches the optimal performance within a few packets, which is much better than a random selection and behaves very similarly in all realizations.

We also demonstrate the performance of the proposed algorithm in the presence of alien interference for LTE channels in Fig. 3. In this scenario, we consider four interferers that use four out of $K = 10$ available channels. These interfering nodes are randomly located outside the network disk and



(a) Sum-Regret.

(b) Spectral Efficiency.

Fig. 3. Performance evaluation over LTE fading channel with alien interference. Simulation parameters: $N = K = 10$, explore length= 500 OFDM symbols, and auction length = 500 OFDM symbols.

within a distance of 500 m from the annular region of the disk. It can be seen from Fig. 3b that the spectral efficiency is reduced by ~2 bits/sec/Hz. However, the proposed algorithm achieves the optimal performance within few thousand symbols similar to the interference-free case, as shown in Fig. 2. This scenario again confirms that our performance is identical to that of the E3 algorithm [19].

VI. Conclusions

In this paper, we presented a novel distributed algorithm for channel allocation over wireless channels where links initially have no estimation for the statistics of the channels. Learning the statistics of the channels in real-time (exploration) comes at the expense of using the best known channels (exploitation). The scenario is modeled as a multiplayer multi-armed bandit problem where a collision occurs if two or more links transmit on the same channel. We proved that our algorithm achieves the optimal order of regret — $\mathcal{O}(\log T)$. Our algorithm is based on a distributed auction algorithm that uses CSMA to avoid the need for an auctioneer (base station). In contrast to the state-of-the-art algorithms, our algorithm requires neither centralized management nor any communication between devices, which makes it very relevant to cognitive ad-hoc networks. From a practical point of view, this results in a significant complexity reduction of the physical layer design.

Acknowledgment

This research was supported by the ISF-NRF Joint research Program, under grant ISF 2277/16 and by ISF 1644/18. This work was supported in part by Singapore NRF2015-NRF-ISF001-2277. S. M. Zafaruddin was partially funded by the Israeli Planning and Budget Committee (PBC) post-doctoral fellowship (2016-2018).

References

[1] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 3, no. 2, pp. 10–31, Second 2000.

[2] M. E. Tanab and W. Hamouda, "Resource allocation for underlay cognitive radio networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1249–1276, Secondquarter 2017.

[3] H. Kwon, S. Kim, and B. G. Lee, "Opportunistic multi-channel CSMA protocol for OFDMA systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, pp. 1552–1557, May 2010.

[4] Y. Yaffe, A. Leshem, and E. Zehavi, "Stable matching for channel access control in cognitive radio systems," in *2010 2nd International Workshop on Cognitive Information Processing*, June 2010, pp. 470–475.

[5] 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, January 2012.

[6] O. Naparstek and A. Leshem, "Bounds on the expected optimal channel assignment in rayleigh channels," in *2012 IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2012, pp. 294–298.

[7] D. P. Bertsekas, "The auction algorithm: A distributed relaxation method for the assignment problem," *Annals of Operations Research*, vol. 14, no. 1, pp. 105–123, Dec 1988. [Online]. Available: https://doi.org/10.1007/BF02186476

[8] K. Yang, N. Prasad, and X. Wang, "An auction approach to resource allocation in uplink OFDMA systems," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4482–4496, Nov 2009.

[9] M. Bayati, B. Prabhakar, D. Shah, and M. Sharma, "Iterative scheduling algorithms," in *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, May 2007, pp. 445–453.

[10] M. Bayati, D. Shah, and M. Sharma, "Max-product for maximum weight matching: Convergence, correctness, and LP duality," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1241–1251, March 2008.

[11] 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, Jan 2014.

[12] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, April 2011.

[13] K. Liu and Q. Zhao, "Cooperative game in dynamic spectrum access with unknown model and imperfect sensing," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1596–1604, April 2012.

[14] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, Nov 2010.

[15] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, Oct 2013.

[16] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits–a musical chairs approach," in *International Conference on Machine Learning*, 2016, pp. 155–163.

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

[18] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[19] N. Nayyar, D. Kalathil, and R. Jain, "On regret-optimal learning in decentralized multi-player multi-armed bandits," *IEEE Transactions on Control of Network Systems*, vol. PP, no. 99, pp. 1–1, 2016.

[20] I. Bistritz and A. Leshem, "Distributed multi-player bandits - a game of thrones approach," in *Advances in Neural Information Processing Systems 31*.

[21] I. Bistritz and A. Leshem, "Game of Thrones: Fully Distributed Learning for Multi-Player Bandits," *arXiv e-prints*, p. arXiv:1810.11162, Oct 2018.

[22] S. M. Zafaruddin, I. Bistritz, A. Leshem, and D. Niyato, "Distributed Learning for Channel Allocation Over a Shared Spectrum," *IEEE JSAC issue on Machine Learning in Wireless Communications (under review), arXiv e-prints*, p. arXiv:1902.06353, Feb 2019.

[23] O. Naparstek and A. Leshem, "A fast matching algorithm for asymptotically optimal distributed channel assignment," in *2013 18th International Conf. on Digital Signal Proc. (DSP)*, July 2013, pp. 1–6.

[24] H. Kuhn, "The hungarian method for the assignment problem," *Naval Res. Logist. Quarter*, vol. 2, no. 1-2, pp. 83–97, 1955.