

Distributed Learning over Markovian Fading Channels for Stable Spectrum Access

Tomer Gafni and Kobi Cohen

Abstract— We consider the problem of multi-user spectrum access in wireless networks. The bandwidth is divided into K orthogonal channels, and M users aim to access the spectrum. Each user chooses a single channel for transmission at each time slot. The state of each channel is modeled by a restless unknown Markovian process. Previous studies have analyzed a special case of this setting, in which each channel yields the same expected rate for all users. By contrast, we consider a more general and practical model, where each channel yields a different expected rate for each user. This model adds a significant challenge of how to efficiently learn a channel allocation in a distributed manner to yield a global system-wide objective. We adopt the stable matching utility as the system objective, which is known to yield strong performance in multichannel wireless networks, and develop a novel Distributed Stable Strategy Learning (DSSL) algorithm to achieve the objective. We prove theoretically that DSSL converges to the stable matching allocation, and the regret, defined as the loss in total rate with respect to the stable matching solution, has a logarithmic order with time. Finally, simulation results demonstrate the strong performance of the DSSL algorithm.

I. INTRODUCTION

We consider the spectrum access problem, where a shared bandwidth is divided into K orthogonal channels (i.e., sub-bands), and M users want to access the spectrum, where $K \geq M$. Each channel is modeled by a Finite-State Markovian Channel (FSMC), which is independent and non-identically distributed across channels. The FSMC is a tractable model widely used to capture the time-varying behavior of a radio communication channel [2], [3]. It is often employed to model radio channel dynamics due to primary user occupancy effects in hierarchical cognitive radio networks (where the M secondary (unlicensed) users are cognitive in terms of learning and adapting good access strategies), or the external interference effects in the open sharing model among M users in the wireless network (e.g., ISM band) [4], [5]. At each time step, each user experiences a different transmission rate over each channel depending on its FSMC distribution, where the FSMC parameters (i.e., the transition probabilities that govern the Markov chain) are unknown. At each time step, each user is allowed to choose one channel to access, and observe the instantaneous channel state. If two users or more access the same channel at the same time, a collision occurs and the achievable rate is zero.

We adopt the stable matching utility (see Section II for details) as the system objective, which is known to yield strong

performance in multichannel wireless networks [6]. We define the *regret* as the loss in total rate with respect to the stable matching solution with known FSMCs. The objective is to develop a distributed learning algorithm for channel allocation and access under unknown FSMCs that minimizes the growth rate of the regret with time t .

A. Main Results

The stable matching problem for multi-user spectrum access was first introduced in [6] under the assumption that the expected rates are known, and a distributed opportunistic CSMA algorithm that solves the problem was proposed. The model with an unknown expected rate matrix and rested setting (i.e., the states of the Markovian process do not change if not observed by the user) was studied in [7], [8]. A regret (with respect to the optimal allocation) of near- $O(\log t)$ was achieved. However, these algorithms require intensive communication between users in order to apply the auction algorithm [9]. In [10], the authors reduced the communication burden, but without guarantees on the achievable regret. Recently, it was shown in [11], [12] that achieving a sum-regret of near- $O(\log t)$ is possible without communication between users, but only for the case of i.i.d channels. In this paper we focus on the general case where the channel states may change whether or not they are being observed (i.e., the restless Markovian setting), and improve the regret scaling with the system parameters by a simple distributed implementation. The main contributions are summarized below.

a) *A general model for spectrum access using a restless Markovian channel model:* As explained above, by contrast to [6]–[8], [10]–[12], in this paper we first solve the channel allocation and access problem under general unknown restless Markovian channel model. Handling this model adds significant challenges in algorithm design and regret analysis. Due to the restless nature of the channels and potential reward loss due to transient effects as compared to steady state when switching channels, learning the Markovian channel characteristics requires that the channels be accessed in a judicious consecutive manner for a period of time. This is reflected in a novel algorithm design that guarantees efficient learning, as detailed next.

b) *Algorithm Development:* We are facing an online learning problem constituted by the well-known exploration versus exploitation dilemma. To remedy this, we propose a novel Distributed Stable Strategy Learning (DSSL) algorithm for solving the problem. Since the FSMCs are unknown, the rate means must be learned by accessing all channels via exploration phases. This results in increasing the regret, since the stable allocation is not performed. Thus, the exploration time must be minimized, while guaranteeing efficient learning.

Tomer Gafni and Kobi Cohen are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501 Israel. Email: gafnito@post.bgu.ac.il, yakovsec@bgu.ac.il.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

A short version of this paper was presented at the 57th Annual Allerton Conference on Communication, Control, and Computing, 2019 [1].

Roughly speaking, each channel can be learned by different exploration times, depending on its unknown parameters (see more details in Section III-D). The algorithm design in this paper contributes to both tackling the more general model, as well as improving the learning efficiency in a fully-distributed manner. Specifically, in existing algorithms [7], [8], [10]–[12], the exploration phase of all channels is determined by the channel that requires the largest exploration time. This results in oversampling the channels and significantly increases the regret. By contrast, the DSSL algorithm estimates online the desired (unknown) exploration rate of each channel. Thus, by sampling the channels according to the desired exploration rate, it avoids oversampling the channels, and thus reduces the regret scaling significantly as compared to existing algorithms.

c) Performance analysis: In terms of theoretical performance analysis, we prove that the DSSL algorithm converges to the stable matching allocation, and the regret has a logarithmic order with time. When comparing to existing approaches, DSSL achieves this under the more general restless Markovian model, and also has significantly better scaling with the system parameters. Specifically, under a common benchmark setting of equal rates among users (but still vary among channels), and $K > M$, which allows a theoretical comparison of learning efficiency between different algorithms, in [8] and [13] the regret scales as $O(\frac{MK}{(\Delta_{\min})^2} \log(t))$, in [12] as $O(\frac{M^3K}{(\Delta_{\min})^2} \log(t))$ and in [11] the regret scales as $O(\frac{MK^2}{(\Delta_{\min})^2} \log(t))$, where Δ_{\min} is the difference in rates between the M th and $(M + 1)$ th best channels. In contrast, under DSSL, the regret scales as $O(\frac{1}{(\Delta_{\min})^2} + MK) \log(t)$. In addition, extensive numerical experiments were performed to demonstrate the efficiency of the proposed DSSL algorithm.

B. Related Work

A number of studies have developed distributed learning algorithms for a special case of the restless Markovian channel model considered in this paper, where each channel yields the same expected rate for all users [14]–[16]. This special case significantly simplifies the channel allocation problem and the analysis (for instance, switching between assigned users does not affect the resulting regret in this special case). In this paper, we consider the general model where each channel yields a different expected rate for each user. This models the situation of different channel fading states across users and channels in actual wireless networks, and adds a significant challenge of how to learn the desired channel allocation in a distributed manner to achieve a global system-wide objective.

Another set of related work on multi-user channel allocation has approached it from the angle of game theoretic and congestion control ([17]–[27] and references therein), hidden channel states [28], and graph coloring ([29]–[32] and references therein). The game theoretic aspects of the problem have been investigated from both non-cooperative (i.e., each user aims at maximizing an individual utility) [18], [19], [24], [25], [33], and cooperative (i.e., each user aims at maximizing a system-wide global utility) [17], [26], [34], [35] settings. Model-free learning strategies were developed in [36], [37] for orthogonal channels, compact models [38], and multiple access channel strategies were developed in [39], [40]. Graph coloring formulations have dealt with modeling

the spectrum access problem as a graph coloring problem, in which users and channels are represented by vertices and colors, respectively (see [29]–[32] and references therein for related studies). Finally, none of these studies have considered the problem of achieving provable stable strategies in the learning context under unknown restless Markovian dynamics, as considered in this paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a wireless network consisting of K orthogonal channels indexed by the set $\mathcal{K} = \{1, 2, \dots, K\}$ and M cognitive users (referred to as users) indexed by the set $\mathcal{M} = \{1, 2, \dots, M\}$, where $K \geq M$. The users aim at accessing the spectrum to send their data. Each user is allowed to choose a single channel for transmission at each time slot, and transmit if the channel is not occupied by a primary user. The users operate in a synchronous time-slotted fashion. Due to spatial geographic dispersion, each user can potentially experience different achievable rates over the channels. When a user i transmits on channel k (when the channel is free) at time slot t , its data rate is given by $r_{i,k}(t)$. This information is concisely represented by an $M \times K$ rate matrix $V(t) = \{r_{i,k}(t)\}$, $i = 1, \dots, M, k = 1, \dots, K$.

We consider the case where the rate process $r_{i,k}(t)$ is Markovian and has a well-defined steady state distribution. The transition probabilities associated with the Markov chain are unknown to the users. The process $r_{i,k}(t)$ evolves independently of the user's actions (i.e., external process). Furthermore, the channel states may change depending on whether or not they are observed (i.e., restless setting). Specifically, the rate of user i on channel k , $r_{i,k}(t)$, is modeled as a discrete time, irreducible and aperiodic Markov chain on a finite-state space $\mathcal{X}^{i,k}$ and is represented by a transition probability matrix $P^{i,k} \triangleq (p_{x,x'}^{i,k} : x, x' \in \mathcal{X}^{i,k})$. The process mean (i.e., the expected rate) is denoted by $\mu_{i,k}$ and is unknown to the users. We define the $M \times K$ expected rate matrix by $U = \{\mu_{i,k}\}$, $i = 1, \dots, M, k = 1, \dots, K$.

Let $X_{i,k}(t)$ be the actual achievable rate for user i on channel k at time t . If two or more users choose to access the same channel at the same time slot, a collision occurs. In this case, $X_{i,k}(t) = 0$. Otherwise, if user i has accessed channel k without colliding with other users, then $X_{i,k}(t) = r_{i,k}(t)$. The users implement carrier sensing to observe the current channel state at each time slot as is typically done in cognitive radio networks [14], [22]. Hence, the channel states are observed regardless of collisions. The transmission scheme for the multi-user spectrum access model is detailed in Section III.

A. Notations

We present the other notations that are used throughout the paper. Let $\bar{\pi}_{i,k} \triangleq (\pi_{i,k}^x, x \in \mathcal{X}^{i,k})$ be the stationary distribution of the Markov chain $P^{i,k}$, and let:

$$\pi_{\min} \triangleq \min_{i \in \mathcal{M}, k \in \mathcal{K}, x \in \mathcal{X}^{i,k}} \pi_{i,k}^x, \quad \hat{\pi}_{i,k}^x \triangleq \max\{\pi_{i,k}^x, 1 - \pi_{i,k}^x\}, \quad \hat{\pi}_{\max} \triangleq \max_{i \in \mathcal{M}, k \in \mathcal{K}, x \in \mathcal{X}^{i,k}} \{\pi_{i,k}^x, 1 - \pi_{i,k}^x\}.$$

We define $X_{\max} \triangleq \max_{i \in \mathcal{M}, k \in \mathcal{K}} \{|\mathcal{X}^{i,k}|\}$ as the maximal cardinality among the state spaces, and

$$x_{\max} \triangleq \max_{i \in \mathcal{M}, k \in \mathcal{K}, x \in \mathcal{X}^{i,k}} x, \quad r_{\max} \triangleq \max_{i \in \mathcal{M}, k \in \mathcal{K}} \sum_{x \in \mathcal{X}^{i,k}} x.$$

Let $\lambda_{i,k}$ be the second largest eigenvalue of $P^{i,k}$, and $\lambda_{\max} \triangleq \max_{i \in \mathcal{M}, k \in \mathcal{K}} \lambda_{i,k}$ be the maximal one among all channels and users. Also, $\bar{\lambda}_{\min} \triangleq 1 - \lambda_{\max}$, $\bar{\lambda}_{i,k} \triangleq 1 - \lambda_{i,k}$ is the eigenvalue gap. Let $M_{x,y}^{i,k}$ be the mean hitting time of state y starting at initial state x for channel k used by user i , and $M_{\max}^{i,k} \triangleq \max_{x,y \in \mathcal{X}^{i,k}, x \neq y} M_{x,y}^{i,k}$. We also define:

$$A_{\max} \triangleq \max_{i,k} (\min_{x \in \mathcal{X}^{i,k}} \pi_{i,k}^x)^{-1} \sum_{x \in \mathcal{X}^{i,k}} x,$$

and

$$L \triangleq \frac{28x_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}{\lambda_{\min}}. \quad (1)$$

The expectations $\mu_{i,k}$ are given by:

$$\mu_{i,k} = \sum_{x \in \mathcal{X}^{i,k}} x \cdot \pi_{i,k}^x,$$

and we define σ_i , for $i = 1, \dots, M$, as a permutation of $\{1, \dots, K\}$ such that

$$\mu_{i,\sigma_i(1)} > \mu_{i,\sigma_i(2)} > \dots > \mu_{i,\sigma_i(K)}.$$

B. A Stable Channel Allocation

Let $a_i(t) \in \mathcal{K}$ be a selection rule, indicating which channel is chosen by user i at time t , which is a mapping from the observed history of the process (i.e., all past actions and observations up to time $t-1$) to $\{1, \dots, K\}$. The expected aggregated data rate for all users up to time t is given by:

$$R(t) = \mathbb{E} \left[\sum_{n=1}^t \sum_{i=1}^M X_{i,a_i(n)}(n) \right]. \quad (2)$$

A policy ϕ_i is a time series vector of selection rules: $\phi_i = (a_i(t), t = 1, 2, \dots)$ for user i .

Definition 1 ([6]): A bipartite matching between channels and users is a permutation $P: \mathcal{M} \rightarrow \mathcal{K}$. The optimal centralized allocation problem is to find a bipartite matching:

$$\mathbf{k}^{**} = \arg \max_{\mathbf{k} \in \mathcal{P}} \sum_{i=1}^M \mu_{i,k(i)}.$$

Definition 2 ([6]): A matching $S: \mathcal{M} \rightarrow \mathcal{K}$ is stable if for every $i \in \mathcal{M}$ and $k \in \mathcal{K}$ satisfying $S(i) \neq k$, if $\mu_{i,S(i)} < \mu_{i,k}$ then there exists some user $i' \in \mathcal{M}$ such that $S(i') = k$ and $\mu_{i',k} > \mu_{i,k}$.

Achieving the optimal allocation in Definition 1 requires implementing a centralized solution, or a distributed solution with heavy complexity and slow convergence [41]. Therefore, we are interested in developing a distributed algorithm with low complexity that converges to the stable matching solution in Definition 2 which is known to yield strong performance and very fast convergence (when the expected rates are known) by using distributed opportunistic CSMA (see Section III-B and [6] for more details on opportunistic CSMA for stable channel allocation).

We assume that the entries in the matrix U are all different, as in [6], which holds in wireless networks due to continuous-valued Shannon rates¹. Thus, there is a unique stable matching

solution under our assumptions, and the expected aggregated rate under the stable matching solution S is given by:

$\sum_{i=1}^M \mu_{i,S(i)}$. The channel $S(i)$ (i.e., the channel that user i selects under the stable matching configuration) is referred to as the *stable channel selection* of user i .

Remark 1: We point out that under an i.i.d. or rested² Markovian channel model, the optimal policy is to transmit on the same channels that achieves the optimal centralized allocation in terms of the sum expected rate. However, the optimal policy in the restless Markovian setting has been shown to be P-SPACE hard even under known Markovian dynamics [42]. Therefore, a commonly adopted approach in this setting is to use a weaker definition of the regret, first introduced in [43] and used later; e.g., in [14], [15], [44], [45], where the policy is compared to a "partially informed" genie who knows the expected rates of the channels, instead of the complete system dynamics. In this paper we adopt this approach as well.

C. The Objective

Since the expected rates $\mu_{i,k}$ are unknown in our setting, the users must learn this information online effectively so as to converge to the stable matching solution. A widely used performance measure of online learning algorithms is the regret, which is defined as the reward loss with respect to an algorithm with a side information on the model. In our setting, we define the regret for policy $\phi = (\phi_i, 1 \leq i \leq M)$ as the loss in the expected aggregated data rate with respect to the stable matching solution that uses the true expected rates:

$$r_{\phi}(t) \triangleq t \cdot \sum_{i=1}^M \mu_{i,S(i)} - \mathbb{E}_{\phi} \left[\sum_{n=1}^t \sum_{i=1}^M X_{i,\phi_i(n)}(n) \right]. \quad (3)$$

A policy ϕ that achieves a sublinear scaling rate of the regret with time (and consequently the time averaged regret tends to zero) approaches the required stable matching solution. The essence of the problem is thus to design an algorithm that learns the unknown expected rates efficiently to achieve the best sublinear scaling of the regret with time.

III. THE DISTRIBUTED STABLE STRATEGY LEARNING (DSSL) ALGORITHM

To achieve the objective, as detailed in Section II-C, we divide the time horizon into three phases, we term exploration, allocation, and exploitation. These three phases are performed repeatedly during the algorithm according to judiciously designed policy rules, as detailed later.

The purpose of the exploration phase is to allow each user to explore all the channels to identify its M best channels (i.e., the M channels that yield the highest expected rates for the user). The users use the sample means as estimators for the expected rates of the channels to achieve this goal. This phase results in a regret loss, since users access sub-optimal channels to explore them, and the stable allocation is not performed. However, this phase is essential to identifying the M best channels and consequently minimizing the regret scaling with

²In the rested model the Markov chain $P^{i,k}$ makes a state transition only when user i accesses channel k .

¹Otherwise, we can add noise to the matrix.

time. The purpose of the exploitation phase is to use the currently learned information to execute the stable matching solution. The allocation phase allows users to allocate the channels among themselves properly in a distributed manner using opportunistic carrier sensing [46].

Since the rate process $r_{i,k}(t)$ can evolve even when channel k is not selected by user i , learning the Markovian rate statistics requires using the channels in a consecutive manner for a period of time [14], [15]. Moreover, frequent switching between channels can cause a loss due to the transient effect. The high-level structure of the DSSL algorithm works as follows. Each user i computes its sufficient number of samples in the exploration phases (condition (13) defined in III-E) for each channel k at the end of every exploitation phase t . If the number of samples is greater than the required number for all k , user i performs another exploitation phase. Otherwise, if the number of samples is smaller than the sufficient number for one or more channels, user i carries out an exploration phase for those channels. When no exploration phase is needed, an allocation phase is performed. At the end of the allocation phase, each user identifies its stable channel selection, and an exploitation phase is carried out. We now discuss the structure of the DSSL algorithm in details.

A. The structure of the exploration phase:

Let $n_O^{i,k}(t)$ be the number of exploration phases in which channel k was selected by user i up to time t . Each exploration phase is divided into two sub epochs: a Random size Epoch (RE), and a Deterministic size Epoch (DE). Let $\gamma^{i,k}(n_O^{i,k}(t) - 1)$ be the last channel state observed at the $(n_O^{i,k}(t) - 1)^{th}$ exploration phase. RE starts at the beginning of the exploration phase until state $\gamma^{i,k}(n_O^{i,k}(t) - 1)$ is observed. This epoch ensures that the generated sample path (after removing the samples observed in the RE epochs) is equivalent to a sample path generated by continuously sensing the Markovian channel without switching. This step guarantees a consistent estimation of the expected rates. Then, DE starts by sensing the channel for a deterministic period of time $4^{n_O^{i,k}(t)}$. The deterministic period of time grows geometrically with time to ensure a relatively small number of channel switching.

B. The structure of the allocation phase:

The allocation phase applies opportunistic CSMA among users. In opportunistic CSMA, the backoff function maps from an index (i.e., expected rate) to a backoff time [46]. The backoff function decreases monotonically with the rates, so that the user with the highest rate on a certain channel waits the minimal time before transmission. All other users sense that the channel is occupied and do not transmit on that channel. To obtain the stable matching allocation, this procedure continues until all M users occupy M channels. For more details on opportunistic CSMA for stable matching see [6].

The allocation phase has two goals in our setting. The first is to assign channels to users to yield a stable matching solution as in [6]. However, since the expected rates are unknown in our setting, the allocation phase is executed by using the sample means. The second goal is to use the backoff function to identify the differences in sample means among users and channels, which is needed for setting efficient learning rates.

This requires a new mechanism that performs opportunistic CSMA, as detailed below.

Let \mathcal{T}_k be the set of all users that attempt to transmit on channel k at a certain stage of the allocation phase. We initialize the phase by declaring each user to be *unassigned*. We divide the time horizon of the allocation phase into two sub-phases. In the first sub-phase, referred to as S_1 , we perform opportunistic CSMA for stable matching as in [6], while replacing the expected rates by the sample means. Specifically, each unassigned user attempts to transmit on its best channel, out of those it has not yet attempted using opportunistic CSMA. On each channel k , the best user out of \mathcal{T}_k in this sub-phase (S_1) is declared to be assigned. All the other users in \mathcal{T}_k store the sample mean of the assigned user (by mapping from the sensed backoff time to the sample mean). This sub-phase continues until all M users are assigned to M channels. The second sub-phase, referred to as S_2 , is used to obtain the side information required for efficient learning. Specifically, the opportunistic CSMA is executed again, but the assigned users of each channel do not transmit. All other users that attempted to transmit in S_1 transmit again on the same channel k . The sample mean of the best user in S_2 (i.e., the second best user in \mathcal{T}_k for each channel k) is stored by the assigned user. This sub-phase continues until all M users in S_2 were observed, and the phase ends.

An example for $M = K = 3$ is given next. The expected rate matrix is shown in Table I. Table II shows the transmission attempts made by the users in the allocation phase before the stable matching was achieved (the assigned users are shown in bold). At time $t = 1$, each user transmits on its best channel (sub-phase S_1). Users 1 and 2 aim to access the same channel (channel 2), and the channel is assigned to user 2 since it has a higher expected rate on this channel (i.e., smaller backoff time). At time $t = 2$, sub-phase S_2 is performed, in which user 1 transmits again on channel 2. At time $t = 3$, user 1 (the only unassigned user) tries to access its second best channel; i.e., channel 1. However, the channel is assigned to user 3 since it has a higher expected rate. The algorithm continues until the three users are assigned to the three channels.

TABLE I: expected rate matrix

U	channel 1	channel 2	channel 3
user 1	45	70	35
user 2	30	90	60
user 3	65	10	50

TABLE II: allocation phase

Sub-phase	Time	channel 1	channel 2	channel 3
S_1	t=1	3	1,2	
S_2	t=2		1	
S_1	t=3	1,3	2	
S_2	t=4	1		
S_1	t=5	3	2	1

C. The structure of the exploitation phase:

Let $n_I(t)$ be the number of exploitation phases up to time t . In the exploitation phase, each user transmits on the channel it was assigned according to the last allocation phase (during

S_1) for a deterministic period of time $2 \cdot 4^{n_i(t)-1}$ (for the n_i^{th} exploitation phase). There are no channel switching and no sample mean updating during the exploitation phase.

D. Parameter setting for efficient learning:

As discussed earlier, exploring the channels increases the regret since the stable matching allocation is not used. On the other hand, it is essential to reduce the estimation error and hence reduce the regret scaling order with time. In this section, we establish the sufficient exploration rate of each channel for each user to achieve efficient learning of the stable matching allocation. We next establish two parameters used in the learning strategy.

1) *Identifying M best channels:* We show in the analysis that a user (say user i) who is interested in distinguishing with a sufficiently high accuracy between two channels k, l that yield expected rates $\mu_{i,k}, \mu_{i,l}$, respectively, must explore them at least $\frac{4L}{(\mu_{i,k} - \mu_{i,l})^2} \cdot \log(t)$ times. Let \mathcal{M}_i be the set of the M best channels of user i . For each channel $k \in \mathcal{M}_i$ we define the deterministic row³ exploration coefficient as

$$D_{i,k}^{(R)} \triangleq \frac{4L}{\min_{\ell \neq k} \{(\mu_{i,k} - \mu_{i,\ell})^2\}}, \quad (4)$$

and for channel $k \notin \mathcal{M}_i$,

$$D_{i,k}^{(R)} \triangleq \frac{4L}{(\mu_{i,k} - \mu_{i,\sigma_i(M)})^2}. \quad (5)$$

Since the expected rates are unknown, the users need to estimate $D_{i,k}^{(R)}$ for each channel $k \in \mathcal{K}$. This estimator is denoted by $\widehat{D}_{i,k}^{(R)}(t)$. Let $\bar{s}_{i,k}(t)$ be the mean transmission rate of user i on channel k . Thus, the adaptive row exploration coefficient for channels $k \in \mathcal{M}_i$ is defined by

$$\widehat{D}_{i,k}^{(R)}(t) \triangleq \frac{4L}{\max\{\Delta_{\min}^2, \min_{\ell \neq k} \{(\bar{s}_{i,k}(t) - \bar{s}_{i,\ell}(t))^2\} - \epsilon\}}, \quad (6)$$

and similarly for $k \notin \mathcal{M}_i$ we have:

$$\widehat{D}_{i,k}^{(R)}(t) \triangleq \frac{4L}{\max\{\Delta_{\min}^2, (\bar{s}_{i,k}(t) - \bar{s}_{i,\sigma_i(M)}(t))^2 - \epsilon\}}, \quad (7)$$

where Δ_{\min} is the smallest difference between two entries in the expected rate matrix U ; i.e.,

$$\Delta_{\min} \triangleq \min_{i \in \mathcal{M}} \Delta_i, \\ \Delta_i \triangleq \min_{k, \ell \in \mathcal{K}, k \neq \ell} |\mu_{i,k} - \mu_{i,\ell}|.$$

2) *CSMA protocol identification:* Consistent with the opportunistic CSMA protocol described above, each user i needs to distinguish between a channel $k \in \mathcal{T}_k$ (this channel is in \mathcal{M}_i as well), and the best channel in \mathcal{T}_k (and the second best channel in \mathcal{T}_k if k is the best channel in \mathcal{T}_k), for all k . Hence, we define the deterministic column exploration coefficient for user i for channel $k \in \mathcal{T}_k$ by:

$$D_{i,k}^{(C)} \triangleq \frac{4L}{(\mu_{i,k} - \max_{j \neq i, j \in \mathcal{T}_k} \mu_{j,k})^2}, \quad (8)$$

³This definition is consistent with the definition of the $M \times K$ expected rate matrix by $U = \{\mu_{i,k}\}$, $i = 1, \dots, M, k = 1, \dots, K$.

and the adaptive column exploration coefficient by:

$$\widehat{D}_{i,k}^{(C)}(t) \triangleq \frac{4L}{\max\{\Delta_{\min}^2, (\bar{s}_{i,k}(t) - \max_{j \neq i} \bar{s}_{j,k}(t))^2 - \epsilon\}}. \quad (9)$$

Note that $\max_{j \neq i, j \in \mathcal{T}_k} \bar{s}_{j,k}(t)$ is known to user i by the design of the opportunistic CSMA (by sub-phase S_2). By combining (4) and (8), the deterministic exploration-rate coefficient of user i for channels $k \in \mathcal{M}_i \cap \mathcal{T}_k$ is given by:

$$D_{i,k} \triangleq \max\{D_{i,k}^{(R)}, D_{i,k}^{(C)}\}, \quad (10)$$

and by combining (6) and (9), the adaptive exploration-rate coefficient of user i for channels $k \in \mathcal{M}_i \cap \mathcal{T}_k$ is given by:

$$\widehat{D}_{i,k}(t) = \max\{\widehat{D}_{i,k}^{(R)}(t), \widehat{D}_{i,k}^{(C)}(t)\}. \quad (11)$$

Remark 2: The design of the adaptive exploration-rate coefficients under DSSL significantly reduces the regret as compared to existing algorithms that use deterministic exploration-rate coefficients determined by the channel that requires the largest exploration time [8], [10]–[12]. For example, consider the expected rate matrix U given in Table I, where parameter L in (1) equals 10^4 . In Table III, we present the deterministic exploration-rate coefficients $D_{i,k}$ defined in (10) for each channel-user pair under DSSL, where $D_{i,k} \cdot \log(t)$ is the number of samples required to achieve consistent estimates of the expected rates. By contrast, in other existing algorithms [8], [10]–[12], all channels are explored with the same exploration-rate coefficient, which is inversely proportional to the squared difference between the mean rate of the optimal allocation and the second best one. When applying this to our example, each channel should be explored for $1600 \cdot \log(t)$ time steps (as seen in Table IV), which significantly increases the exploration times unnecessarily, and consequently increases the regret.

TABLE III: Exploration coefficients under the DSSL algorithm

$D_{i,k}$	channel 1	channel 2	channel 3
user 1	400	100	400
user 2	45	100	45
user 3	178	25	178

TABLE IV: Exploration coefficients under other existing algorithms [8], [10]–[12]

$D_{i,k}$	channel 1	channel 2	channel 3
user 1	1600	1600	1600
user 2	1600	1600	1600
user 3	1600	1600	1600

E. Choosing between phases types:

Since $D_{i,k}$ is unknown, the algorithm replaces $D_{i,k}$ by its estimate $\widehat{D}_{i,k}(t)$. Furthermore, to ensure that $\widehat{D}_{i,k}(t)$ overestimates $D_{i,k}$, the users need to sense at least $I \cdot \log(t)$ times each of their channels in exploration phases, where

$$I \triangleq \frac{7\epsilon^2}{48(r_{\max} + 2)^2 \cdot L}, \quad (12)$$

which can be viewed as the rate function of the estimators among all channels. At the end of the exploitation phases, the

users check the condition:

$$T_{i,k}^{(O)}(t) > \max \left\{ \widehat{D}_{i,k}(t), \frac{2}{I} \right\} \cdot \log(t), \quad (13)$$

where $T_{i,k}^{(O)}(t)$ is the number of samples in the exploration phases accessed in sub epochs DE for user i on channel k up to time t .

If the condition holds for user i , the user enters another exploitation phase by transmitting on the same channel in which it transmitted during the last exploitation phase. Otherwise, if the condition does not hold, the user enters an exploration phase by sensing channel k . At the end of the phase, the user signals the other users that it has finished the exploration phase. If such an interruption occurred, all the users again check condition (13). If it holds for all users, they start an allocation phase. At the end of the allocation phase, an exploitation phase starts. A pseudocode of the DSSL algorithm is provided in Algorithm 1.

Algorithm 1 DSSL Algorithm for user i

Initialization: For all K channels, execute an exploration phase where a single observation is taken from each channel;
while $t \leq T$ **do**
 if Condition (13) does not hold for channel k **then**
 Enter an exploration phase with length $4n_{O}^{i,k}(t)$;
 Update $\bar{s}_{i,k}(t)$ and increment $n_{O}^{i,k}(t) = n_{O}^{i,k}(t) + 1$;
 goto step 3
 end if
 Send an interrupt signal;
 Start an allocation phase;
 Start an exploitation phase with length $2 \cdot 4^{n_I(t)}$. If an interruption occurs, go to step 3;
 $n_I(t) = n_I(t) + 1$;
end while

IV. REGRET ANALYSIS

Success in obtaining a logarithmic regret order depends on how fast $\widehat{D}_{i,k}(t)$ converges to a value which is no smaller than $D_{i,k}$ (so that user i senses channel k at least $D_{i,k} \cdot \log t$ time slots in most of the times). The analysis in the Appendix shows that exploring channels as in (13) guarantees the desired convergence speed. Specifically, in the following theorem we establish a finite-sample bound on the regret with time, which results in a logarithmic scaling of the regret.

Theorem 1: Assume that the proposed DSSL algorithm is implemented and that the assumptions on the system model described in Section II hold. Then, the regret at time t is upper bounded by:

$$\begin{aligned} r(t) &\leq A_{\max} \cdot \left(\sum_{i=1}^M \sum_{k=1}^K (\lceil \log_4(3A_{i,k} \log(t) + 1) \rceil + 1) \right) \\ &+ \sum_{i=1}^M \sum_{k=1}^K \left[\left(4A_{i,k} \cdot \log(t) + 1 \right. \right. \\ &\quad \left. \left. + M_{\max}^{i,k} (\lceil \log_4(3A_{i,k} \log(t) + 1) \rceil + 1) \right) \right. \\ &\quad \left. \cdot \left(\mu_{i,S(i)} + \mu_{S^{-1}(k),k} - \mu_{i,k} \right) \right] \end{aligned}$$

$$\begin{aligned} &+ M^2 \cdot A_{\max} \cdot \left(\sum_{i=1}^M \sum_{k=1}^K (\lceil \log_4(3A_{i,k} \log(t) + 1) \rceil + 1) \right) \\ &+ \left[\left(2e \log(M+1) \right) \right. \\ &\quad \left. \cdot \left(\sum_{i=1}^M \sum_{k=1}^K (\lceil \log_4(3A_{i,k} \log(t) + 1) \rceil + 1) \right) \right] \\ &\cdot \left[\sum_{j=1}^M \mu_{j,S(j)} \right] \\ &+ \left(A_{\max} + (M^2K + MK) \frac{6X_{\max}}{\pi_{\min}} \left(\sum_{j=1}^M \mu_{j,S(j)} \right) \right) \\ &\cdot \left(\lceil \log_4\left(\frac{3}{2}t + 1\right) \rceil \right) + O(1), \end{aligned} \quad (14)$$

where $A_{i,k}$ is given by:

$$A_{i,k} \triangleq \begin{cases} \max\{2/I, D_{i,k}^{(\max)}\}, & \text{if } k \in \mathcal{G}_i \\ \max\{2/I, 4L/\Delta_{\min}^2\}, & \text{if } k \notin \mathcal{G}_i \end{cases}, \quad (15)$$

\mathcal{G}_i is defined as the set of all indices $k \in \mathcal{K}$ of user i that satisfy:

$$\min_{\ell \neq k} \{(\mu_{i,k} - \mu_{i,\ell})^2, (\mu_{i,k} - \max_{j \neq i} \mu_{j,k})^2\} - 2\epsilon > \Delta_{\min}^2,$$

for $k \in \mathcal{T}_k$, and

$$(\min_{\ell \neq k} \{(\mu_{i,k} - \mu_{i,\ell})^2\} - 2\epsilon > \Delta_{\min}^2,$$

for $k \notin \mathcal{T}_k$, where $D_{i,k}^{(\max)}$ is defined as:

$$D_{i,k}^{(\max)} \triangleq \frac{4L}{\min_{\ell \neq k} \{(\min_{\ell \neq k} \{(\mu_{i,k} - \mu_{i,\ell})^2, (\mu_{i,k} - \max_{j \neq i} \mu_{j,k})^2\} - 2\epsilon\}}. \quad (16)$$

The proof is given in the Appendix.

Note that Theorem 1 shows that similar to [8], [11]–[13], the regret under DSSL has a logarithmic order with time. DSSL, however, achieves this under the more general restless Markovian model, and also has significantly better scaling with M, K and Δ_{\min} . Specifically, under a common benchmark setting of equal rates among users (but still vary among channels), and $K > M$, which allows a theoretical comparison of learning efficiency between different algorithms, in [8] and [13] the regret scales as $O(\frac{MK}{(\Delta_{\min})^2} \log(t))$, in [12] as $O(\frac{M^3K}{(\Delta_{\min})^2} \log(t))$ and in [11] the regret scales as $O(\frac{MK^2}{(\Delta_{\min})^2} \log(t))$. In contrast, under DSSL, the regret scales as $O(\frac{1}{(\Delta_{\min})^2} + MK) \log(t)$ due to the novel algorithm design that explores every channel according to its unique adaptive exploration rate, while guaranteeing efficient learning.

V. SIMULATION RESULTS

In this section we present simulation results to evaluate the performance of DSSL numerically. In Subsection V-A we start by evaluating the convergence of DSSL under unknown restless fading FSMCs with respect to the stable matching solution solved under known restless fading FSMCs. We also evaluate the performance as compared to random allocation and the optimal centralized allocation schemes. Then, in

Section V-B we examine the learning efficiency of DSSL as compared to other online learning algorithms under unknown restless FSMC, and verify our theoretical logarithmic regret. We performed 1,000 Monte-Carlo experiments and averaged the performance over the experiments.

A. Convergence of DSSL to stable matching

We start by describing the wireless channel model used in the simulations. Each user experiences a block fading channel which remains constant during each time slot, and varies between time slots. The channel response experienced by user i at time slot t is given by $h(i, t) = r(i, t)e^{j\rho(i, t)}$, where $r(i, t) = |h(i, t)|$ denotes the channel rate, and $\rho(i, t)$ denotes the channel phase experienced by user i at time t . Let $f(i, r)$ denote the Probability Density Function (PDF) of the fading channel rate $r(i)$ experienced by user i (e.g., Rayleigh fading distribution in the simulations). We consider independent but non-identically distributed channels across users, and Markovian correlated channels across time slots. The FSMC model [2], [3] partitions the range of the channel gain values into a finite number of intervals and represents each interval as a state of a Markov chain. The thresholds of the intervals at user i are denoted by $\tau_n(i), n = 0, \dots, N$, where $0 = \tau_0(i) < \tau_1(i) < \dots < \tau_{N-1}(i) < \tau_N(i) = \infty$. The channel rate $r(i, t)$ experienced by user i is said to be in state $g_n(i), 1 < n < N$, if it lies in the interval: $t_{n-1}(i) \leq r(i, t) < \tau_n(i)$. The states are partitioned to yield an equal initial state probability for all states:

$$\int_{\tau_{n-1}(i)}^{\tau_n(i)} f(i, r) dr = \frac{1}{N}, n = 1, \dots, N.$$

The transition probability to transition from state $g_n(i)$ to state $g_\ell(i)$ is defined by:

$$p_{n,\ell}(i) \triangleq \Pr(\tau_{\ell-1}(i) \leq r(i, t+1) < \tau_\ell(i) \mid \tau_{n-1}(i) \leq r(i, t) < \tau_n(i))$$

where $r(i, t)$ and $r(i, t+1)$ are the current channel gain and the channel gain in the next time slot experienced by user i , respectively. In the simulations, we quantized the channel gain to 6 states; i.e., $N = 6$, and we simulated a case of 3 users and 5 channels. The transition probability matrix P and the expected rate matrix U are given by:

$$P = \begin{pmatrix} 3/6 & 2/6 & 1/6 & 0 & 0 & 0 \\ 2/8 & 3/8 & 2/8 & 1/8 & 0 & 0 \\ 1/9 & 2/9 & 3/9 & 2/9 & 1/9 & 0 \\ 0 & 1/9 & 2/9 & 3/9 & 2/9 & 1/9 \\ 0 & 0 & 1/8 & 2/8 & 3/8 & 2/8 \\ 0 & 0 & 0 & 1/6 & 2/6 & 3/6 \end{pmatrix},$$

$$U = \begin{pmatrix} 45 & 70 & 35 & 17.5 & 12.5 \\ 27.5 & 90 & 60 & 15 & 20 \\ 65 & 10 & 50 & 16.5 & 30 \end{pmatrix}.$$

We compared the expected rate evolution of DSSL under unknown FSMCs against stable matching, random allocation and the optimal centralized allocation solved under known FSMCs. The optimal centralized algorithm served as an upper bound benchmark for all algorithms, and the stable matching served as an upper bound for DSSL. In the random allocation scheme

users access an arbitrary channel with equal probability. As shown in Fig. 1 the average rate under DSSL converged to that of the stable matching, as desired. The stable matching allocation allocates user 1 to channel 3, user 2 to channel 2, and user 3 to channel 1. Fig. 2 shows that the average achievable rate of each user in the DSSL algorithm converged to the stable allocation.

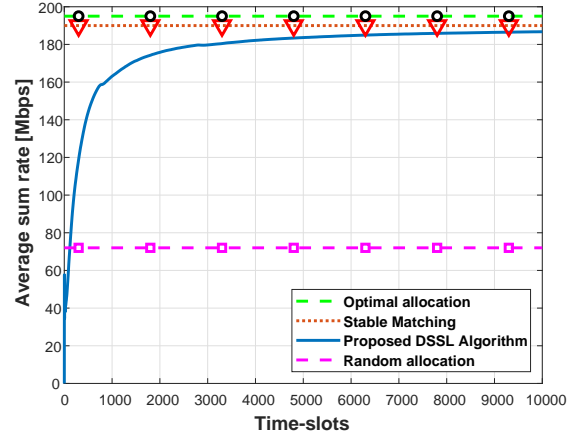


Fig. 1: Comparison of the system average rate of various schemes

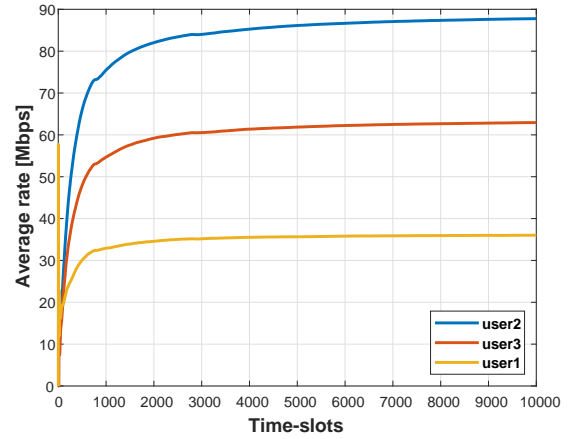


Fig. 2: Comparison of users' average rate for the proposed DSSL algorithm

B. Learning efficiency of DSSL

We next evaluated the learning efficiency of DSSL as compared to other online learning algorithms under unknown restless FSMCs. We considered the hierarchical access channel model in spectrum access networks. This models the situation of primary and secondary users that share the spectrum. Primary users (licensed) occupy the spectrum occasionally, and a secondary user is allowed to transmit over a single channel when the channel is free. Thus, each channel has two states, *good* (free) and *bad* (occupied). The good state results in a positive expected rate, whereas bad state result in a zero rate. The occupancies of the channels by the primary users are modeled as Markov processes (i.e., Gilbert-Elliott channel).

First, we simulated a special case of our model where each channel yielded the same expected rate for all users. In [14],

[15], the RCA and DSEE algorithms were proposed to solve this special case. The RCA algorithm performs random regenerative cycles until catching predefined states in each phase, which results in oversampling the channels, and therefore is expected to increase the regret as compared to DSSL. The DSEE algorithm overcomes this issue by performing deterministic sequencing for both the exploration and exploitation phases. However, the deterministic sequencing requires the algorithm to explore all channels using the maximal exploration rate among all channels, which is expected to increase the regret as compared to DSSL (that learns the desired exploration rate for each channel) as well. We simulated the case of 2 users, 6 channels, each with two states: 0, 1. The transition probabilities for all channels to transition from 0 to 1 and from 1 to 0, respectively, were $p_{01} = [0.1, 0.1, 0.5, 0.1, 0.1, 0.7]$, $p_{10} = [0.2, 0.3, 0.1, 0.4, 0.5, 0.08]$, the expected rates for all channels at states 1, 0, respectively, are $r_1 = [1, 1, 1, 1, 1, 1]$, $r_0 = [0.1, 0.1, 0.1, 0.1, 0.1, 0.1]$. As can be seen in Fig. 3, the DSSL algorithm outperformed both RCA and DSEE and achieved the logarithmic regret order with time.

Finally, we simulated the scenario where the stable matching allocation was also the optimal centralized allocation, and the channels were i.i.d. across time slots (and not Markovian). We compared DSSL to the dE^3 algorithm which was designed for this setting. However, dE^3 requires communication between users since it implements a distributed auction that requires users to observe the bids of other users [8]. We used the same parameters as selected and tuned by the authors in [8]. Similar to the DSEE algorithm, in dE^3 the exploration-rate coefficient was determined by the channel with the largest exploration time. Thus, we expected that DSSL would yield a faster convergence rate due to the adaptive design of the exploration epochs. As shown in Fig. 4, DSSL indeed outperformed the dE^3 algorithm.

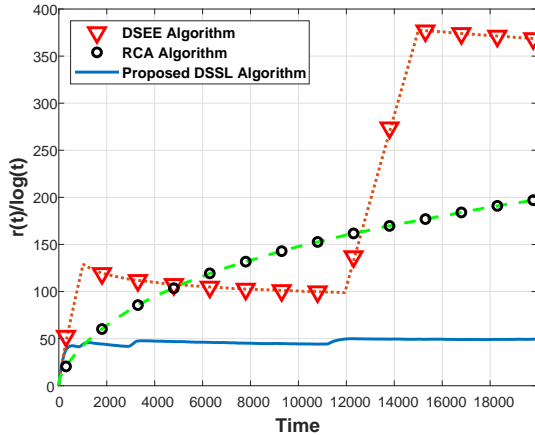


Fig. 3: The regret (normalized by $\log t$) under DSSL, DSEE, and RCA as a function of time. Parameter setting: 2 users, 6 channels, each with two states: 0, 1. Transition probabilities for all channels to transition from 0 to 1 and from 1 to 0, respectively: $p_{01} = [0.1, 0.1, 0.5, 0.1, 0.1, 0.7]$, $p_{10} = [0.2, 0.3, 0.1, 0.4, 0.5, 0.08]$, expected rates for all channels at states 1, 0, respectively: $r_1 = [1, 1, 1, 1, 1, 1]$, $r_0 = [0.1, 0.1, 0.1, 0.1, 0.1, 0.1]$.

VI. CONCLUSION

We developed a novel algorithm for the multi-user spectrum access problem in wireless networks, dubbed the Distributed

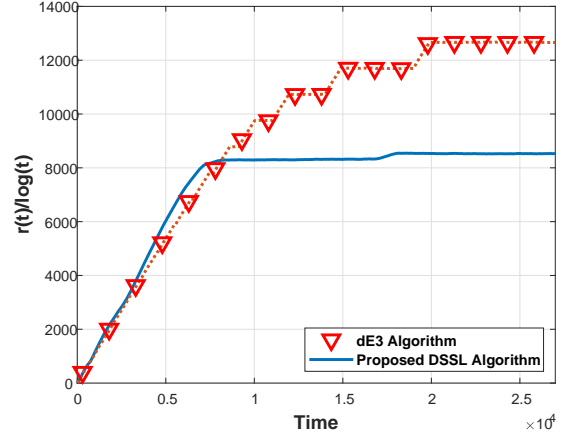


Fig. 4: The regret under DSSL and dE^3 as a function of time. Parameter setting: 3 users, 3 channels, with mean transmission rates: $[0.2, 0.25, 0.3; 0.4, 0.6, 0.5; 0.7, 0.9, 0.8]$.

Stable Strategy Learning (DSSL) algorithm. In contrast to existing models, for the first time we considered the case of restless Markov channels, which requires a different algorithm structure to accurately learn the channel statistics. Moreover, the channels selection rules are adaptive in order to reduce the exploration time required for efficient learning. We showed theoretically that DSSL achieves a logarithmic regret with time, and better regret scaling with the system parameters as compared to existing approaches that have studied special cases of the model. Extensive simulation results supported the theoretical study and demonstrated the strong performance of DSSL.

VII. APPENDIX

In this appendix we prove Theorem 1.

Definition 1: Let T_1 be the smallest integer, such that for all $t \geq T_1$ the following holds: $D_{i,k} \leq \hat{D}_{i,k}(t)$ for all $i \in \mathcal{M}, k \in \mathcal{K}$, and also $\hat{D}_{i,k}(t) \leq D_{i,k}^{(\max)}$ for all $i \in \mathcal{M}, k \in \mathcal{G}_i$.

Lemma 1: Assume that the DSSL algorithm is implemented as described in Section III. Then, $E(T_1) < \infty$ is bounded independent of t .

Proof: $E(T_1)$ can be written as follows:

$$\begin{aligned} E[T_1] &= \sum_{n=1}^{\infty} n \cdot \Pr(T_1 = n) = \sum_{n=1}^{\infty} \Pr(T_1 \geq n) \\ &= \sum_{n=1}^{\infty} \Pr\left(\bigcup_{i \in \mathcal{M}} \bigcup_{k \in \mathcal{G}_i} \bigcup_{l=n}^{\infty} (\hat{D}_{i,k}(l) < D_{i,k} \text{ or } \hat{D}_{i,k}(l) > D_{i,k}^{(\max)}) \text{ or } \right. \\ &\quad \left. \bigcup_{i \in \mathcal{M}} \bigcup_{k \notin \mathcal{G}_i} \bigcup_{l=n}^{\infty} (\hat{D}_{i,k}(l) < D_{i,k})\right) \\ &\leq \sum_{i \in \mathcal{M}} \sum_{k \in \mathcal{G}_i} \sum_{n=1}^{\infty} \sum_{l=n}^{\infty} \Pr(\hat{D}_{i,k}(l) < D_{i,k} \text{ or } \hat{D}_{i,k}(l) > D_{i,k}^{(\max)}) \\ &\quad + \sum_{i \in \mathcal{M}} \sum_{k \notin \mathcal{G}_i} \sum_{n=1}^{\infty} \sum_{l=n}^{\infty} \Pr(\hat{D}_{i,k}(l) < D_{i,k}) \end{aligned}$$

Note that if we show that

$$\Pr(\hat{D}_{i,k}(l) < D_{i,k} \text{ or } \hat{D}_{i,k}(l) > D_{i,k}^{(\max)}) \leq C \cdot l^{-(2+\delta)} \quad (17)$$

for some constants $C > 0, \delta > 0$ for all $i \in \mathcal{M}, k \in \mathcal{G}_i$ for all $l \geq n$, then we get:

$$\sum_{i \in \mathcal{M}} \sum_{k \in \mathcal{G}_i} \sum_{n=1}^{\infty} \sum_{l=n}^{\infty} \Pr(\hat{D}_{i,k}(l) < D_{i,k} \text{ or } \hat{D}_{i,k}(l) > D_{i,k}^{(\max)})$$

$$\begin{aligned}
&\leq MK \cdot C \left[\sum_{l=1}^{\infty} l^{-(2+\delta)} + \sum_{n=2}^{\infty} \sum_{l=n}^{\infty} l^{-(2+\delta)} \right] \\
&\leq MK \cdot C \left[\sum_{l=1}^{\infty} l^{-(2+\delta)} + \sum_{n=2}^{\infty} \int_{n-1}^{\infty} l^{-(2+\delta)} dl \right] \\
&= MK \cdot C \left[\sum_{l=1}^{\infty} l^{-(2+\delta)} + \frac{1}{1+\delta} \sum_{n=2}^{\infty} (n-1)^{-(1+\delta)} \right] < \infty,
\end{aligned}$$

which is bounded independent of t . Similarly, showing that $\Pr(\widehat{D}_{i,k}(l) < D_{i,k}) \leq C \cdot l^{-(2+\delta)}$ for some constants $C, \delta > 0$ for all $i \in \mathcal{M}, k \notin \mathcal{G}_i$ for all $j \geq n$ completes the statement. We start bounding (17). We look at the first inequality of (17) for user i with channel $k \in \mathcal{M}_i \cap \mathcal{T}_k$. The event $\widehat{D}_{i,k}(t) < D_{i,k}$ implies:

$$\begin{aligned}
&\max \left\{ \Delta_{\min}^2, \min \left\{ \min_{\ell \neq k} \{(\bar{s}_{i,k}(t) - \bar{s}_{i,\ell}(t))^2\} - \epsilon, \right. \right. \\
&\quad \left. \left. (\bar{s}_{i,k}(t) - \max_{j \neq i} \bar{s}_{j,k}(t))^2 - \epsilon \right\} \right\} \\
&> \min \left\{ \min_{\ell \neq k} \{(\mu_{i,k} - \mu_{i,\ell})^2\}, (\mu_{i,k} - \max_{j \neq i} \mu_{j,k})^2 \right\},
\end{aligned}$$

which after algebraic manipulations implies that at least one of the following holds:

$$\begin{aligned}
&\min_{\ell \neq k} \{(\bar{s}_{i,k}(t) - \bar{s}_{i,\ell}(t))^2\} - \epsilon > \min_{\ell \neq k} \{(\mu_{i,k} - \mu_{i,\ell})^2\} \\
&(\bar{s}_{i,k}(t) - \max_{j \neq i} \bar{s}_{j,k}(t))^2 - \epsilon > (\mu_{i,k} - \max_{j \neq i} \mu_{j,k})^2.
\end{aligned}$$

Similarly, the second inequality of (17) implies one of the following:

$$\begin{aligned}
&\min_{\ell \neq k} \{(\bar{s}_{i,k}(t) - \bar{s}_{i,\ell}(t))^2\} - \epsilon < \min_{\ell \neq k} \{(\mu_{i,k} - \mu_{i,\ell})^2\} - 2\epsilon \\
&(\bar{s}_{i,k}(t) - \max_{j \neq i} \bar{s}_{j,k}(t))^2 - \epsilon < (\mu_{i,k} - \max_{j \neq i} \mu_{j,k})^2 - 2\epsilon.
\end{aligned}$$

Let $k^* = \arg \min_{\ell \neq k} (\mu_{i,k} - \mu_{i,\ell})^2$ (i.e., $(\mu_{i,k} - \mu_{i,k^*})^2 = \min_{\ell \neq k} (\mu_{i,k} - \mu_{i,\ell})^2$). Cascading the events written above we get :

$$\begin{aligned}
&\Pr(\widehat{D}_{i,k}(t) < D_{i,k} \text{ or } \widehat{D}_{i,k}(t) > D_{i,k}^{(\max)}) \\
&\leq \Pr(|(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t))^2 - (\mu_{i,k} - \mu_{i,k^*})^2| > \epsilon) \\
&+ \Pr(|(\bar{s}_{i,k}(t) - \max_{j \neq i} \bar{s}_{j,k}(t))^2 - (\mu_{i,k} - \max_{j \neq i} \mu_{j,k})^2| > \epsilon).
\end{aligned} \tag{18}$$

Each of the terms in (18) is the probability of a deviation of the squared difference for two Markov chains' sample means from the squared difference of their expected means by an ϵ . We look at the first term of (18). Using conventional steps from set theory, it can be shown that:

$$\begin{aligned}
&\Pr(|(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t))^2 - (\mu_{i,k} - \mu_{i,k^*})^2| > \epsilon) \\
&\leq \left[\Pr(|(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t))[(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t)) \right. \\
&\quad \left. - (\mu_{i,k} - \mu_{i,k^*})]| > \frac{\epsilon}{2}) \right] \\
&+ \left[\Pr(|(\mu_{i,k} - \mu_{i,k^*})[(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t)) \right. \\
&\quad \left. - (\mu_{i,k} - \mu_{i,k^*})]| > \frac{\epsilon}{2}) \right]
\end{aligned}$$

$$\begin{aligned}
&\leq \left[\Pr(|(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t)) - (\mu_{i,k} - \mu_{i,k^*})| > 1) \right. \\
&\quad \left. + \Pr(|(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t)) - (\mu_{i,k} - \mu_{i,k^*})| > \frac{\epsilon}{2(R+1)}) \right] \\
&\quad + \Pr(|(\mu_{i,k} - \mu_{i,k^*}) + 1| > R) \\
&\quad + \left[\Pr(\mu_{i,k} > R') \right. \\
&\quad \left. + \Pr(|(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t)) - (\mu_{i,k} - \mu_{i,k^*})| > \frac{\epsilon}{2(R'+1)}) \right],
\end{aligned}$$

for every $R, R' > 0$. We choose $R = R' = r_{\max} + 1$, hence the third and fourth terms are equal to 0, and we get the concentration inequalities:

$$\Pr(|(\bar{s}_{i,k}(t) - \bar{s}_{i,k^*}(t))^2 - (\mu_{i,k} - \mu_{i,k^*})^2| > \epsilon)$$

$$< 6 \cdot \max \left\{ \Pr(|\bar{s}_{i,k}(t) - \mu_{i,k}| > \frac{\epsilon}{4(r_{\max} + 2)}), \right. \tag{19}$$

$$\left. \Pr(|\bar{s}_{i,k^*}(t) - \mu_{i,k^*}| > \frac{\epsilon}{4(r_{\max} + 2)}) \right\}. \tag{20}$$

Similar bounds can be obtained for the second term in (18). To bound (19) and (20) we use Lezaud's results [47]:

Lemma 2 ([47]): Consider a finite-state, irreducible Markov chain $\{X_t\}_{t \geq 1}$ with state space S , matrix of transition probabilities P , an initial distribution q , and stationary distribution π . Let $N_{\mathbf{q}} = \left\| \left(\frac{q^{(x)}}{\pi^{(x)}}, x \in S \right) \right\|_2$. Let $\widehat{P} = P'P$ be the multiplicative symmetrization of P where P' is the adjoint of P on $l_2(\pi)$. Let $\epsilon = 1 - \lambda_2$, where λ_2 is the second largest eigenvalue of the matrix P' . ϵ will be referred to as the eigenvalue gap of P' . Let $f : S \rightarrow \mathcal{R}$ be such that $\sum_{y \in S} \pi_y f(y) = 0$, $\|f\|_2 \leq 1$ and $0 \leq \|f\|_2^2 \leq 1$ if P' is irreducible. Then, for any positive integer n and

$$\text{all } 0 < \lambda \leq 1, \text{ we have: } P \left(\frac{\sum_{t=1}^n f(X_t)}{n} \geq \lambda \right) \leq N_{\mathbf{q}} \exp$$

$$\left[-\frac{n\lambda^2\epsilon}{12} \right].$$

Consider an initial distribution $\mathbf{q}^{i,k}$ for channel k of user i . We have:

$$N_{\mathbf{q}}^{(i,k)} = \left\| \left(\frac{q_{i,k}^x}{\pi_{i,k}^x}, x \in X^{i,k} \right) \right\|_2 \leq \sum_{x \in X^{i,k}} \left\| \frac{q_{i,k}^x}{\pi_{i,k}^x} \right\|_2 \leq \frac{1}{\pi_{\min}}.$$

We point out that the sample rate mean $\bar{s}_{i,k}(t)$ is computed by $T_{i,k}^{(O)}(t)$ observation taken only from sub epochs DE in the exploration phases, thus the sample path that generated $\bar{s}_{i,k}(t)$ can be viewed as a sample path generated by a Markov chain with a transition matrix identical to the original channel $\{i, k\}$, so we can apply Lezaud's result to bound (19) and (20). For equation (19):

we define $n_x^{i,k}(t)$ to be the number of occurrences of state x on channel k sensed by user i up to time t .

$$\begin{aligned}
&\Pr(\bar{s}_{i,k}(t) - \mu_{i,k} > \frac{\epsilon}{4(r_{\max} + 2)}) \\
&= \Pr \left(\sum_{x \in \mathcal{X}^{i,k}} x \cdot n_x^{i,k}(t) - T_{i,k}^{(O)}(t) \sum_{x \in \mathcal{X}^{i,k}} x \cdot \pi_{i,k}^x > \frac{T_{i,k}^{(O)}(t) \cdot \epsilon}{4(r_{\max} + 2)} \right) \\
&= \Pr \left(\sum_{x \in \mathcal{X}^{i,k}} (x \cdot n_x^{i,k}(t) - T_{i,k}^{(O)}(t) x \cdot \pi_{i,k}^x) > \frac{T_{i,k}^{(O)}(t) \cdot \epsilon}{4(r_{\max} + 2)} \right)
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{x \in \mathcal{X}^{i,k}} \Pr \left(x \cdot n_x^{i,k}(t) - T_{i,k}^{(O)}(t)x \cdot \pi_{i,k}^x > \frac{T_{i,k}^{(O)}(t) \cdot \epsilon}{4(r_{\max}+2)|\mathcal{X}^{i,k}|} \right) \\
&= \sum_{x \in \mathcal{X}^{i,k}} \Pr \left(n_x^{i,k}(t) - T_{i,k}^{(O)}(t) \cdot \pi_{i,k}^x > \frac{T_{i,k}^{(O)}(t) \cdot \epsilon}{4(r_{\max}+2)|\mathcal{X}^{i,k}| \cdot x} \right) \\
&= \sum_{x \in \mathcal{X}^{i,k}} \Pr \left(\frac{\sum_{n=1}^t \mathbf{1}(x_{i,k}(n)=x) - T_{i,k}^{(O)}(t)\pi_{i,k}^x}{\hat{\pi}_{i,k}^x \cdot T_{i,k}^{(O)}(t)} > \frac{T_{i,k}^{(O)}(t) \cdot \epsilon}{4(r_{\max}+2)|\mathcal{X}^{i,k}| \cdot x \hat{\pi}_{i,k}^x} \right) \\
&\leq |\mathcal{X}^{i,k}| \cdot N_{\mathbf{q}}^{(i,k)} \exp \left(-T_{i,k}^{(O)}(t) \cdot \frac{\epsilon^2}{16(r_{\max}+2)^2 \cdot x^2 \cdot |\mathcal{X}^{i,k}|^2 \cdot (\hat{\pi}_{i,k}^x)^2} \cdot \frac{(1-\lambda_{i,k})}{12} \right),
\end{aligned}$$

and from (13), we have: $T_{i,k}^{(O)}(t) > \frac{2}{I} \log(t)$ with I defined in (12). Thus,

$$\Pr \left(|\bar{s}_{i,k}(t) - \mu_{i,k}| > \frac{\epsilon}{4(r_{\max}+2)} \right) \leq \frac{|X_{\max}|}{\pi_{\min}} \cdot t^{-2+\delta}. \quad (21)$$

The same bound can be obtained for (20), and with the same steps, for all terms in (18). The proof for all $i \in \mathcal{M}, k \notin \mathcal{G}_i$ is similar, and thus Lemma 1 follows. \square

We now bound the expected regret defined in (3). We divide the time horizon for $t < T_1$ and $t > T_1$. Since T_1 is finite (due to Lemma 1), the regret for all $t < T_1$ results in a constant term $O(1)$ which is independent of t . For $t > T_1$, we know that the adaptive exploration coefficient is no smaller than the deterministic exploration coefficient, and no larger than $D_{i,k}^{(\max)}$ defined in (16); i.e.,

$$D_{i,k} \leq \hat{D}_{i,k}(t) \leq D_{i,k}^{(\max)}, \quad (22)$$

for all $i \in \mathcal{M}, k \in \mathcal{G}_i$, and the LHS of the inequality for $i \in \mathcal{M}, k \in \mathcal{K}$. Thus, the exploration phases provides sufficient learning for the channel statistics (and the upper bound ensures that the channels are judiciously oversampled in the exploration phases).

We continue bounding the regret for $t > T_1$:

$$r(t) \leq (t - T_1) \cdot \sum_{i=1}^M \mu_{i,S(i)} - \mathbb{E} \left[\sum_{n=T_1+1}^t \sum_{i=1}^M X_{i,a_i(n)}(n) \right]. \quad (23)$$

For convenience, we will develop (23) between $n = 1$ and t with (22) (and the LHS for $k \notin \mathcal{G}_i$) holds for all $1 \leq n \leq t$, which upper bounds (23):

$$\begin{aligned}
r(t) &\leq (t - T_1) \cdot \sum_{i=1}^M \mu_{i,S(i)} - \mathbb{E} \left[\sum_{n=T_1+1}^t \sum_{i=1}^M X_{i,a_i(n)}(n) \right] \\
&\leq t \cdot \sum_{i=1}^M \mu_{i,S(i)} - \mathbb{E} \left[\sum_{n=1}^t \sum_{i=1}^M X_{i,a_i(n)}(n) \right]. \quad (24)
\end{aligned}$$

We can rewrite (24) as:

$$r(t) \leq \sum_{i=1}^M \sum_{k=1}^K (\mu_{i,k} \cdot E[T_{i,k}(t)] - E[\sum_{n=1}^t X_{i,k}(n)]) \quad (25)$$

$$+ \left(t \cdot \sum_{i=1}^M \mu_{i,S(i)} - \sum_{i=1}^M \sum_{k=1}^K \mu_{i,k} \cdot E[T_{i,k}(t)] \right), \quad (26)$$

where $T_{i,k}(t)$ is the total number of transmission for user i on channel k up to time t (and $X_{i,k}(n) = 0$ if user i did not try to access channel k at time n).

Equation (25) can be considered as the regret due to the transient effect (the initial state of the channel may not be given by the stationary distribution), and (26) is the regret caused by not playing the stable matching allocation. Both (25) and (26) can be thought of as the sum of three different regret terms, corresponding to the three phases described in Section III. We denote by $r^O(t), r^A(t), r^I(t)$ the regret caused in the exploration, allocation and exploitation phases respectively; i.e., the regret can be written as:

$$r(t) = r^O(t) + r^A(t) + r^I(t). \quad (27)$$

We next bound the regret in each of the three phases.

Regret in the exploration phases:

To bound the regret in the exploration phases, we first bound the number of exploration phases $n_O^{i,k}(t)$ for each user $i \in \mathcal{M}$ on each channel $k \in \mathcal{K}$ by time t . As described in Section (III-A), the total number of samples from the exploration phases in sub epochs DE for user i on channel k up to time t is:

$$T_{i,k}^{(O)}(t) = \sum_{n=1}^{n_O^{i,k}(t)} 4^{n-1} = \frac{1}{3} (4^{n_O^{i,k}(t)} - 1).$$

Since we are in an exploration phase, from (13) together with (22), we have $T_{i,k}^{(O)}(t) < A_{i,k} \cdot \log(t)$ ($A_{i,k}$ is defined in (15)). Hence,

$$n_O^{i,k}(t) \leq \lceil \log_4(3A_{i,k} \log(t) + 1) \rceil + 1. \quad (28)$$

We use the following lemma to show that the regret caused by channel switching is upper bounded by a constant independent of the number of transmissions on the channel in each phase.

Lemma 3 ([48]): Consider an irreducible, aperiodic Markov chain with state space S , a matrix of transition probabilities P , an initial distribution \vec{q} which is positive in all states, and stationary distribution $\vec{\pi}$ (π_s is the stationary probability of state s). The state (reward) at time t is denoted by $s(t)$. Let μ denote the mean reward. If we play the chain for an arbitrary time T , then there exists a value

$$A_p \leq (\min_{s \in S} \pi_s)^{-1} \sum_{s \in S} s, \text{ such that: } E \left[\sum_{t=1}^T s(t) - \mu T \right] \leq A_p.$$

Lemma 3 bounds the probability of a large deviation from the stationary distribution of a Markov chain (which we refer to as the transient effect). By the construction of the exploration phases described in Section (III-A), in each exploration phase there is no channel switching (each channel has its own unique exploration phases), therefore (25) in the exploration phases is bounded by:

$$A_{\max} \cdot \left(\sum_{i=1}^M \sum_{k=1}^K (\lceil \log_4(3A_{i,k} \log(t) + 1) \rceil + 1) \right). \quad (29)$$

We next bound (26) in the exploration phases. Note that each user has its own exploration time, independent of the other users; i.e., when user i explores, the other users (for which condition (13) holds) continue to exploit. However, user's i exploration may affect other users exploring during that time due to collision. Specifically, when user i explores channel k it affects the regret in two ways. First, user i does not transmit in its stable channel; hence, the regret is increased by $\mu_{i,S(i)} - \mu_{i,k}$. Second, if k is a stable channel of another user, then because of the collision, the regret will increase by $\mu_{S^{-1}(k),k}$ ($S^{-1}(k)$ is the user for which channel k is its stable channel). Combining these two terms, we bound (26) in exploration phases by:

$$\sum_{i=1}^M \sum_{k=1}^K \left(E[N_{i,k}^{(O)}(t)] \cdot (\mu_{i,S(i)} + \mu_{S^{-1}(k),k} - \mu_{i,k}) \right), \quad (30)$$

where $N_{i,k}^{(O)}(t)$ consists of the time indices from RE and DE, and depends on the mean hitting time of the channel due to the regenerative cycles. With (28) we have:

$$\begin{aligned} E[N_{i,k}^{(O)}(t)] &\leq \sum_{n=0}^{n_O^{i,k}-1} (4^n + M_{\max}^{i,k}) \\ &= \frac{1}{3}(4^{n_O^{i,k}}(t) - 1) + M_{\max}^{i,k} \cdot n_O^{i,k}(t) \\ &\leq \frac{1}{3}[4(3A_{i,k} \cdot \log(t) + 1) - 1] \\ &\quad + M_{\max}^{i,k} \cdot \log_4(3A_{i,k} \log(t) + 1). \end{aligned} \quad (31)$$

Combining (29) and (30) we can bound the first term in (27):

$$\begin{aligned} r^O(t) &\leq A_{\max} \cdot \left(\sum_{i=1}^M \sum_{k=1}^K ([\log_4(3A_{i,k} \log(t) + 1)] + 1) \right) \\ &\quad + \sum_{i=1}^M \sum_{k=1}^K \left(E[N_{i,k}^{(O)}(t)] \cdot (\mu_{i,S(i)} + \mu_{S^{-1}(k),k} - \mu_{i,k}) \right), \end{aligned} \quad (32)$$

which coincides with the first and second terms on the RHS of (14).

Regret in the allocation phases:

Since an allocation phase will only come after an exploration phase, the number of allocation phases by time t , $n_A(t)$ is bounded by the total number of exploration phases by time t ; i.e.,

$$n_A(t) \leq \sum_{i=1}^M \sum_{k=1}^K n_O^{i,k}(t),$$

and by using (28) we have:

$$n_A(t) \leq \sum_{i=1}^M \sum_{k=1}^K [\log_4(3A_{i,k} \log(t) + 1)] + 1. \quad (33)$$

Since the expected rates are unknown in our setting, the allocation phase is executed using the sample means. To bound the expected time required for each allocation phase, we use proposition VI.4. in [6]:

Lemma 4 ([6]): Denote the expected delay to reach a stable matching configuration by T_M . There is some constant C s.t. for every M we have:

$$T_M \leq C \log(M + 1).$$

Specifically, it was shown in [6] that it is sufficient to choose $C = 2e$ for the bound to hold.

Lemma 4 states that each allocation phase is finite with respect to t , and only depends on the number of users. The total time in

allocation phases by time t , denoted by $T_A(t)$, can be bounded by combining (33) with lemma 4:

$$\begin{aligned} E[T_A(t)] &\leq (2C \log(M + 1)) \\ &\quad \cdot \left(\sum_{i=1}^M \sum_{k=1}^K [\log_4(3A_{i,k} \log(t) + 1)] + 1 \right), \end{aligned} \quad (34)$$

with $C = 2e$.

We now bound (25) and (26) for the allocation phases. In each allocation phase, the maximum number of channel switchings is $M \cdot M$; thus, the regret caused by the transient effect is bounded by:

$$A_{\max} \cdot M^2 \cdot \left(\sum_{i=1}^M \sum_{k=1}^K ([\log_4(3A_{i,k} \log(t) + 1)] + 1) \right). \quad (35)$$

and the regret due to sub-optimal allocation can be bounded by:

$$E[T_A(t)] \cdot \left(\sum_{i=1}^M \mu_{i,S(i)} \right). \quad (36)$$

Combining (35), (36) we have:

$$\begin{aligned} r^A(t) &\leq A_{\max} \cdot M^2 \cdot \left(\sum_{i=1}^M \sum_{k=1}^K ([\log_4(3A_{i,k} \log(t) + 1)] + 1) \right) \\ &\quad + [(C \log(M + 1)) \cdot \left(\sum_{i=1}^M \sum_{k=1}^K [\log_4(3A_{i,k} \log(t) + 1)] + 1 \right)] \\ &\quad \cdot \left(\sum_{i=1}^M \mu_{i,S(i)} \right), \end{aligned} \quad (37)$$

which coincides with the third and fourth terms in the RHS of (14).

Regret in the exploitation phases:

We first bound the number of exploitation phases up to time t . As described in Section III-C, the number of time slots in the n^{th} exploitation phase is $2 \cdot 4^{(n-1)}$. Thus we have:

$$\sum_{n=1}^{n_I(t)} 2 \cdot 4^{n-1} = \frac{2}{3}(4^{n_I} - 1) \leq t,$$

which implies

$$n_I \leq \lceil \log_4(\frac{3}{2}t + 1) \rceil. \quad (38)$$

During the exploitation phases, there are no channel switchings (each user exploits its stable channel). As a result, the regret caused by the transient effect in the exploitation phases is upper bounded by:

$$A_{\max} \cdot \lceil \log_4(\frac{3}{2}t + 1) \rceil. \quad (39)$$

It remains to bound the regret as a result of not playing the stable matching allocation (which we refer to as a sub-optimal allocation) in the exploitation phases. The event of playing a sub-optimal allocation in an exploitation phase occurs if the previous allocation phase results in a sub-optimal allocation, which occurs if one of the following takes place. The first is that user i did not correctly identify the order of its M best channels entering the allocation phase. This event would be denoted by Y_i . The second eventuality is when the user with the highest expected rate in channel k was not identified

correctly in the allocation phase. This event is denoted by Z_k . We write these events explicitly:

$$Y_i(t_n) = \bigcup_{k \in \mathcal{M}_i} \bigcup_{l \in \mathcal{K}} \{\bar{s}_{i,k}(t_n) < \bar{s}_{i,l}(t_n) | \mu_{i,k} > \mu_{i,l}\}$$

$$Z_k(t_n) = \bigcup_{j \in \mathcal{T}_k} \{\bar{s}_{i,k}(t_n) < \bar{s}_{j,k}(t_n) | \mu_{i,k} = \max_{l \in \mathcal{T}_k} \mu_{l,k}\},$$

where t_n denotes the starting time of the n^{th} exploitation phase. Based on the above notations, the probability for a sub-optimal allocation ($P_S(n)$) in an exploitation phase at time t_n is given by:

$$P_S(n) \triangleq \Pr \left(\bigcup_{i \in \mathcal{M}} Y_i(t_n) \text{ or } \bigcup_{k \in \mathcal{K}} Z_k(t_n) \right).$$

The number of time slots in a sub-optimal allocation in the exploitation phases can be written as:

$$E[\tilde{T}(t)] = \sum_{n=1}^{n_I(t)} 2 \cdot 4^{n-1} \cdot P_S(n) \leq \sum_{n=1}^{\lceil \log_4(\frac{3}{2}t+1) \rceil} 2 \cdot 4^{n-1} \cdot P_S(n)$$

$$\leq \sum_{n=1}^{\lceil \log_4(\frac{3}{2}t+1) \rceil} 3t_n \cdot P_S(n). \quad (40)$$

To complete Theorem 1, we need to show that:

$$P_S(n) = \Pr \left(\bigcup_{i \in \mathcal{M}} Y_i(t_n) \text{ or } \bigcup_{k \in \mathcal{K}} Z_k(t_n) \right) \leq B \cdot t_n^{-1}, \quad (41)$$

for some $B > 0$ (there is only a logarithmic number of terms in (40)). Using union bounds we have:

$$\Pr \left(\bigcup_{i \in \mathcal{M}} Y_i(t_n) \text{ or } \bigcup_{k \in \mathcal{K}} Z_k(t_n) \right) \leq M^2 K \cdot \Pr(\bar{s}_{i,k}(t_n) < \bar{s}_{i,l}(t_n) | \mu_{i,k} > \mu_{i,l}) \quad (42)$$

$$+ M K \cdot \Pr(\bar{s}_{i,k}(t_n) < \bar{s}_{j,k}(t_n) | \mu_{i,k} = \max_{l \in \mathcal{T}_k} \mu_{l,k}) \quad (43)$$

To bound (42) and (43), we define $C_{t,v} = \sqrt{L \log(t)/v}$. Equation (42) implies that at least one of the following must hold

$$\bar{s}_{i,k}(t_n) \leq \mu_{i,k} - C_{t_n, T_{i,k}^{(O)}} \quad (44)$$

$$\bar{s}_{i,l}(t_n) \geq \mu_{i,l} + C_{t_n, T_{i,l}^{(O)}} \quad (45)$$

$$\mu_{i,k} < \mu_{i,l} + C_{t_n, T_{i,l}^{(O)}} + C_{t_n, T_{i,k}^{(O)}}. \quad (46)$$

First we show that the probability for event (46) is zero.

$$\Pr(\mu_{i,k} < \mu_{i,l} + C_{t_n, T_{i,l}^{(O)}} + C_{t_n, T_{i,k}^{(O)}})$$

$$= \Pr \left(\mu_{i,k} - \mu_{i,l} < \sqrt{\frac{L \log t_n}{T_{i,l}^{(O)}(t_n)}} + \sqrt{\frac{L \log t_n}{T_{i,k}^{(O)}(t_n)}} \right)$$

$$\leq \Pr \left(\mu_{i,k} - \mu_{i,l} < 2 \sqrt{\frac{L \log t_n}{\min \{T_{i,k}^{(O)}(t_n), T_{i,l}^{(O)}(t_n)\}}} \right)$$

$$\leq \Pr \left(\min \{T_{i,k}^{(O)}(t_n), T_{i,l}^{(O)}(t_n)\} < \frac{4L}{(\mu_{i,k} - \mu_{i,l})^2} \log(t_n) \right).$$

Combining (22) with (13) (which holds since we started an allocation phase), we have:

$$T_{i,k}^{(O)}(t_n) > \frac{4L}{\min_{\ell \neq k} \{(\mu_{i,k} - \mu_{i,\ell})^2\}} \log(t_n)$$

$$\geq \frac{4L}{(\mu_{i,k} - \mu_{i,l})^2} \log(t_n)$$

$$T_{i,l}^{(O)}(t_n) > \frac{4L}{\min_{j \neq \ell} \{(\mu_{i,l} - \mu_{i,j})^2\}} \log(t_n)$$

$$\geq \frac{4L}{(\mu_{i,k} - \mu_{i,l})^2} \log(t_n),$$

which ensures that the probability of (46) is zero. Note that here we used the fact that $D_{i,k} \geq D_{i,k}^{(R)}$.

We now bound (44) and (45) using Lezaud's result (Lemma 3). With similar steps as used above to bound (19), we can show:

$$\Pr(\bar{s}_{i,k}(t_n) \leq \mu_{i,k} - C_{t_n, v_{i,k}}) \leq \frac{|\mathcal{X}^{i,k}|}{\pi_{\min}} t^{-\frac{L \bar{\lambda}_{\min}}{28 X_{\max}^2 \max^2 \max^2}} \quad (47)$$

$$\Pr(\bar{s}_{i,l}(t_n) \geq \mu_{i,l} + C_{t_n, v_{i,l}}) \leq \frac{|\mathcal{X}^{i,l}|}{\pi_{\min}} t^{-\frac{L \bar{\lambda}_{\min}}{28 X_{\max}^2 \max^2 \max^2}}. \quad (48)$$

Using (1), (42) is bounded by:

$$M^2 K \cdot \Pr(\bar{s}_{i,k}(t_n) < \bar{s}_{i,l}(t_n) | \mu_{i,k} > \mu_{i,l})$$

$$\leq M^2 K \cdot \frac{2X_{\max}}{\pi_{\min}} \cdot t^{-1}. \quad (49)$$

Equation (43) can be bounded using similar techniques, this time using the fact that $D_{i,k} \geq D_{i,k}^{(C)}$, and we can bound (41):

$$\Pr \left(\bigcup_{i \in \mathcal{M}} Y_i(t_n) \text{ or } \bigcup_{k \in \mathcal{K}} Z_k(t_n) \right) \leq (M^2 K + M K) \frac{2X_{\max}}{\pi_{\min}} \cdot t^{-1}. \quad (50)$$

With (50) we can bound (40), and therefore the regret due to sub-optimal allocation in the exploitation phases is bounded by:

$$3 \left(\sum_{i=1}^M \mu_{i, S(i)} \right) (M^2 K + M K) \frac{2X_{\max}}{\pi_{\min}} \cdot \lceil \log_4(\frac{3}{2}t + 1) \rceil. \quad (51)$$

By combining (51) with (39), the total regret in the exploitation phases is:

$$r^I(t) \leq A_{\max} \cdot \lceil \log_4(\frac{3}{2}t + 1) \rceil$$

$$+ 3 \left(\sum_{i=1}^M \mu_{i, S(i)} \right) (M^2 K + M K) \frac{2X_{\max}}{\pi_{\min}} \cdot \lceil \log_4(\frac{3}{2}t + 1) \rceil, \quad (52)$$

which coincides with the two last terms on the RHS of (14).

REFERENCES

- [1] T. Gafni and K. Cohen, "A distributed stable strategy learning algorithm for multi-user dynamic spectrum access," in *57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 347–351, 2019.
- [2] H. S. Wang and N. Moayeri, "Finite-state markov channel-a useful model for radio communication channels," *IEEE transactions on vehicular technology*, vol. 44, no. 1, pp. 163–171, 1995.

- [3] P. Sadeghi, R. A. Kennedy, P. B. Rapajic, and R. Shams, "Finite-state markov modeling of fading channels—a survey of principles and applications," *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 57–80, 2008.
- [4] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 79–89, 2007.
- [5] N. Slamnik-Kriještorac, H. Krešo, M. Ruffini, and J. M. Marquez-Barja, "Sharing distributed and heterogeneous resources toward end-to-end 5g networks: A comprehensive survey and a taxonomy," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1592–1628, 2020.
- [6] 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, 2012.
- [7] 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.
- [8] N. Nayyar, D. Kalathil, and R. Jain, "On regret-optimal learning in decentralized multiplayer multiarmed bandits," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 597–606, 2016.
- [9] 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, 1988.
- [10] O. Avner and S. Mannor, "Multi-user lax communications: a multi-armed bandit approach," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, IEEE, 2016.
- [11] I. Bistriz and A. Leshem, "Distributed multi-player bandits—a game of thrones approach," in *Advances in Neural Information Processing Systems*, pp. 7222–7232, 2018.
- [12] E. Boursier, V. Perchet, E. Kaufmann, and A. Mehrabian, "A Practical Algorithm for Multiplayer Bandits when Arm Means Vary Among Players," *arXiv e-prints*, p. arXiv:1902.01239, Feb 2019.
- [13] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.
- [14] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5588–5611, 2012.
- [15] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2012.
- [16] T. Gafni and K. Cohen, "Learning in restless multi-armed bandits using adaptive arm sequencing rules," in *Proc. of the IEEE International Symposium on Information Theory (ISIT)*, pp. 1206–1210, Jun. 2018.
- [17] Z. Han, Z. Ji, and K. R. Liu, "Fair multiuser channel allocation for OFDMA networks using Nash bargaining solutions and coalitions," *IEEE Transactions on Communications*, vol. 53, no. 8, pp. 1366–1376, 2005.
- [18] I. Menache and N. Shimkin, "Rate-based equilibria in collision channels with fading," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 7, pp. 1070–1077, 2008.
- [19] U. O. Candogan, I. Menache, A. Ozdaglar, and P. A. Parrilo, "Competitive scheduling in wireless collision channels with correlated channel state," in *Game Theory for Networks, 2009. GameNets '09. International Conference on*, pp. 621–630, 2009.
- [20] I. Menache and A. Ozdaglar, "Network games: Theory, models, and dynamics," *Synthesis Lectures on Communication Networks*, vol. 4, no. 1, pp. 1–159, 2011.
- [21] L. M. Law, J. Huang, and M. Liu, "Price of anarchy for congestion games in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3778–3787, 2012.
- [22] K. Cohen, A. Leshem, and E. Zehavi, "Game theoretic aspects of the multi-channel ALOHA protocol in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, pp. 2276–2288, 2013.
- [23] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "Fasa: Accelerated S-ALOHA using access history for event-driven M2M communications," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 6, pp. 1904–1917, 2013.
- [24] C. Singh, A. Kumar, and R. Sundaresan, "Combined base station association and power control in multichannel cellular networks," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 1065–1080, 2016.
- [25] K. Cohen and A. Leshem, "Distributed game-theoretic optimization and management of multichannel aloha networks," *IEEE/ACM Transactions on Networking*, vol. 24, no. 3, pp. 1718–1731, 2016.
- [26] K. Cohen, A. Nedić, and R. Srikant, "Distributed learning algorithms for spectrum sharing in spatial random access wireless networks," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2854–2869, 2017.
- [27] D. Malachi and K. Cohen, "Queue and channel-based aloha algorithm in multichannel wireless networks," *IEEE Wireless Communications Letters*, vol. 9, no. 8, pp. 1309–1313, 2020.
- [28] M. Yemini, A. Leshem, and A. Somekh-Baruch, "Restless hidden markov bandits with linear rewards," *arXiv preprint arXiv:1910.10271*, 2019.
- [29] W. Wang and X. Liu, "List-coloring based channel allocation for open-spectrum wireless network," in *proc. of IEEE Vehic. Tech. Conf.*, 2005.
- [30] J. Wang, Y. Huang, and H. Jiang, "Improved algorithm of spectrum allocation based on graph coloring model in cognitive radio," in *WRI International Conference on Communications and Mobile Computing*, vol. 3, pp. 353–357, 2009.
- [31] A. Checco and D. Leith, "Learning-based constraint satisfaction with sensing restrictions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, pp. 811–820, Oct 2013.
- [32] A. Checco and D. J. Leith, "Fast, responsive decentralised graph colouring," *arXiv preprint arXiv:1405.6987*, 2014.
- [33] H. Cao and J. Cai, "Distributed opportunistic spectrum access in an unknown and dynamic environment: A stochastic learning approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4454–4465, 2018.
- [34] A. Leshem and E. Zehavi, "Bargaining over the interference channel," in *IEEE International Symposium on Information Theory*, pp. 2225–2229, 2006.
- [35] I. Bistriz and A. Leshem, "Approximate best-response dynamics in random interference games," *IEEE Transactions on Automatic Control*, vol. 63, no. 6, pp. 1549–1562, 2018.
- [36] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks," in *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, 2017.
- [37] 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, 2019.
- [38] D. Livne and K. Cohen, "PoPS: Policy Pruning and Shrinking for deep reinforcement learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 4, pp. 789–801, 2020.
- [39] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2897–2911, 2020.
- [40] K. Cohen and D. Malachi, "A time-varying opportunistic multiple access for delay-sensitive inference in wireless sensor networks," *IEEE Access*, vol. 7, pp. 170475–170487, 2019.
- [41] 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.
- [42] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, 1999.
- [43] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.
- [44] A. Lesage-Landry and J. A. Taylor, "The multi-armed bandit with stochastic plays," *IEEE Transactions on Automatic Control*, vol. 63, no. 7, pp. 2280–2286, 2017.
- [45] P. Reverdy, V. Srivastava, and N. E. Leonard, "Satisficing in multi-armed bandit problems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3788–3803, 2016.
- [46] 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. 231–241, 2005.
- [47] P. Lezaud, "Chernoff-type bound for finite markov chains," *Annals of Applied Probability*, pp. 849–867, 1998.
- [48] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 977–982, 1987.