Human-Behavior and QoE-Aware Dynamic Channel Allocation for 5G Networks: A Latent Contextual Bandit Learning Approach

Pan Zhou, *Member, IEEE*, Jie Xu, *Member, IEEE*, Wei Wang, *Senior Member, IEEE*, Changkun Jiang, *Member, IEEE,* Kehao Wang, *Member, IEEE,* Jia Hu, *Member, IEEE*

***Abstract*—With the rapid advance of smart wireless technologies, a plethora of human behavioral data are generated in 5G networks, which is reported capable to improve network performance by leveraging intelligent channel resource allocation through big data analytics. However, what information can be extracted for the network mobility management, how to exploit the knowledge for resource allocation and to meet the user-centric quality of experience (QoE) are not well understood and fully explored. To address this problem, we propose an online learning algorithm for dynamic channel allocation based on contextual multi-armed bandit (CMAB) theory. Especially, we divide the stochastic human behavioral data into two categories: the user location and the QoE-driven context. Noticing that the distributions of CSI vary spatially, we define a set of user's geographic locations that shares the same set of CSI distributions as a *cluster*, and the stochastic channel distributions vary across clusters. The problem is formulated as a novel *latent* SCB problem, where the proposed agnostic SCB algorithm could automatically find the underlying clusters and significantly improve the learning performance. We then extend our online learning algorithm into the practical multi-user random access scenario. We conduct experiments on a real dataset collected from China Mobile, which indicate that our algorithms outperform existing approaches tremendously and perform extremely well in *large-scale* and *high-mobility* networks.**

***Index Terms*—Human behavior, QoE, 5G, Contextual bandits, Channel allocation, User mobility, Online learning.**

I. Introduction

Nowadays, an increasing number of diversified mobile devices are being deployed in hot spots to meet the explosive demand of high-data-rate applications in 5G networks. Along with rapid technology advance, mobile users are preferring to carry a smartphone with several sensors (e.g., GPS measurement and video recording), use Google glasses, wear smart-shirts, smart watches and shoes with sensors to enjoy satisfactory multimedia services. During this process, a plethora of quality of service (QoS) data (e.g., packet delay, bandwidth, throughput and peak signal to noise ratio (PSNR)) and human behavioral data (e.g. location or mobility pattern, factors of satisfaction in quality of experience (QoE)) are generated and collected.

In comparision, existing 4G networks providing all-IP (Internet Protocol) broadband access are based on a *reactive* mechanism, leading to very poor spectrum efficiency. To tackle this problem, artificial intelligence (AI) and its sub-categories like machine learning has been evolving as a golden rule, where nowadays it allows 5G networks to be *predictive* and *proactive*. Hence, this idea is essential in making the 5G vision conceivable. Moreover, as the big data analytics flourishes, much attention has been drawn to exploiting the predicted human behavior data to improve the performance of future wireless systems during next decades, which is referred to as the *context-aware* resource allocations [1]–[3]. However, most of human behavioral data are not well understood and fully explored in the design loop of mobile networks [4].

In the 5G networks, the human behavior data are regarded as context information, which is considered relevant to the interaction between a mobile user and a 5G application. To support context-aware wireless access, the context information are often translated from the measured data that are acquired and stored in the network management side. As noticed, the aforementioned QoS indicators are mainly employed to quantify some network-centric and service performance, rather than to directly meet user demand, which pays more attention to user's personalized satisfactory level of services. In this regard, the quality of experience (QoE) is believed to be a more appropriate criterion, and a paradigm shift from the technological QoS metrics to superior QoE is desirable [6].

Moreover, flourishing and diversified applications in 5G systems are vulnerable to user mobility and uncertainty in dynamics of channel state information (CSI). On the one hand, the mobility of users is not purely random but exhibit significant amount of predictability, which may be learned after monitoring user movements for a period of time. On the other hand, accurately predicting CSI is a significantly key precondition to decide data transmission schemes for wireless communications. The strength attenuation of wireless signal is named as channel fading, which is divided into *large timescale (slow)* fading and *small timescale (fast)* fading. Hence, acquiring small timescale (say in micro-seconds) fading of CSI receives more attention [15], [16] for its great importance to evaluate the wireless communication performance. Moreover, the human behavior and QoE related context information, e.g., PSNR, dropping probability, location and traffic map,

Pan Zhou is with School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China.

Jie Xu is with Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124, USA.

Wei Wang is with College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China.

Changkun Jiang is with College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China.

Kehao Wang is with School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China.

Jia Hu is with Computer Science, the University of Exeter, Exeter, EX4 4QF, UK.

Emails: panzhou@hust.edu.cn, jiexu@miami.edu, wangw@zju.edu.cn, jiangchangkun@gmail.com, kehao.wang@whut.edu.cn, J.Hu@exeter.ac.uk

are usually varies and predictable in a large timescale (say in seconds, minutes, hours or even days). How to correlate the available/predictable information in different time scales together for the optimal channel resource allocation is still open for future wireless communications.

Since no statistical tool exists to model and resolve the above important problem, one promising approach is to devise machine learning-driven policy for actively learning, steering the user behavioral data, and adapting the prediction of small timescale CSI to match the large timescale context information in QoE-aware network performance optimization. This motivates our current work. In this paper, we classify the human behavioral data into two categories: user location and QoE related context. The user location data come along with the user mobility, and users at different locations probably experience different channel distributions due to different communication environmental profiles. We define a set of geographic locations that shares the same set of channel state distributions as a *cluster*, and the stochastic channel distributions vary across clusters. Plus, the QoE-aware contextual data describing the QoE metric are user-specific, where their distributions only differ among different type of users.

Observing the QoE related factors as side contextual information, e.g., PSNR, throughput and blocking probability and video frames rate, our policy as an active learner can predict the CSI and select a communication channel based on its accumulated historic statistics to maximize the QoE performance metric [6]. This problem can be analyzed within the contextual multi-armed bandit (MAB) framework [8], [9], where the QoE-aware contextual information that summarizing the QoE factors could be encoded perfectly as a Stochastic Contextual multi-armed Bandit (SCB) problem [10]. However, the context alone may not be sufficient for mobile communications. When a user is moving around over time, the communication environment and the corresponding channel distributions probably change in different clusters. This phenomenon indicates that the classic assumption of fixed set of channel distributions over space [31]–[37] is restrictive and impractical.

To tackle the more practical problem, we design an agnostic online algorithm to learn the optimal channel allocation strategy and user clusters over time in terms QoE. We consider stochastic user behaviors and channel distributions, which are based on the stochastic model of MAB theory [11]. Determining the underlying clusters from many mobile users' communication locations becomes a very challenging problem, since the clusters are not observable and need to be learned gradually over time. The repeated process of first deciding underlying types (of clusters) and then selecting an arm (channel) out of many forms a latent bandit problem [12] in the machine learning society. As a nutshell, with context-awareness, our study belongs to the *Latent Contextual MAB* (LCMAB) problem. In this case, a reward of QoE is revealed only after given a certain user location and context for an allocated channel. Like any MAB algorithm, the goal is to minimize the expected cumulative *regret* of the policy as reward loss compared to a genie-aided policy that always chooses optimal arm and context for the LCMAB problem.

Furthermore, we consider the multi-user random access (RA) scenario in 5G networks, and extend the medium access control problem in our online learning setting. Unlike classic MAB frameworks [31]–[37] that always seeking the optimal channel with largest channel reward for a single user over time, we consider a more general regret model of selecting the channel with the D-th largest reward to prioritized user, where D=1 stands for the optimal channel. This is motivated by the fact that users with different *payment abilities* or *service types* must have priorities in experiencing different QoEs. For example, a paid user should allocate better channels over that of free users for video services. We investigate and provide the regret performance of our algorithm in multiple users RA scenario under the typical TDMA, CSMA, ALOHA, successive interference cancellation (SIC) protocols. Our contributions are summarized as follows:

1) We first consider a simple and ideal case that a user's mobility is within the same *known* underlying cluster, and the QoE-aware dynamic channel allocation is formulated as a Stochastic Contextual multi-armed Bandit (SCB) problem. We present a novel policy that is referred to as SCB(D), which is a non-trivial generalization of UCB1 in [11] and DCB(ϵ) in [10]. The SCB(D) provides a general solution for selecting a channel with the D-th largest expected rewards under the observed QoE factors as contexts for the SCB problem.
2) Then, we focus on the user's mobility is from different and unknown clusters with completely unknown channel and context distributions, which is formulated as a LCMAB problem. We propose an agnostic algorithm named as A-SCB(D) to adaptively learn the underlying cluster and perform the QoE-aware channel allocation over time. We provide the upper bounds of the regret performance of A-SCB(D). The key idea of A-SCB(D) is to group the user's locations with the same (or similar) context and channel distributions and learn the same underlying clusters to speed up the learning performance.
3) Our algorithm is also extended to the multi-user RA scenario, and we implement all the proposed algorithms on a real LTE downlink dataset collected from China Mobile. Under a typical QoE metric, we have shown that our proposed A-SCB(D) tremendously outperform the SCB(D) and the vanilla UCB1 [11]. This provides strong support that the incorporating human behavioral data indeed has great potential to improve the performance of 5G networks.

The rest of this paper is organized as follows. Section II discusses the works that related to this paper. Section III presents the system model. Section IV focuses on an ideal case that the QoE-aware dynamic channel allocation is formulated as a SCB problem with a single known cluster. Section V consists our main technical innovations and focused on the LCMAB with unknown user mobility. We show that it is a very challenging issue and proposed an agnostic version of SCB(D) in this setting. Section VI extends our theory to the multiple user RA setting. Experiments results are available in Section VII, and the conclusion and future work is drawn in

Section VIII.

II. Related Works

Recently, growing research interest is shown in applying machine learning techniques to wireless communications and networking problems, e.g., CSI prediction by deep learning [17], short-term fading channel prediction by extreme learning [18], traffic-aware online network selection by MAB [19], throughput optimization under unknown interference by online learning [20], QoE prediction by multiple machine learning algorithms [21], cell prediction by supervised learning [23]. etc. Meanwhile, many research works on user mobility prediction are also presented. For example, authors in [6] presented a support vector machine (SVM) scheme to predict next cell using short term CSI and long term handover history information in real time. However, the impact of user mobility and location on QoE for various services is not well understood.

Especially, reinforcement learning as a hot online learning algorithm, is widely applied in the wireless communication and networking problems [25]. Actually, the multi-armed bandit (MAB) problem belongs to the catalogue of reinforcement learning, which both share the same dilemma of the *exploitation* and *exploration* tradeoff [26]. And, the contextual MAB (CMAB) is just a special case of MAB. However, in contrast, the MAB (or CMAB) could provide the anytime optimal solutions rather than providing the asymptotically optimal solution in the reinforcement learning, so it is more applicable for the mission critical applications with relatively short learning periods. This is also true for context-aware reinforcement learning methods, such as [27]. Moreover, previous work had found that, in many traditional reinforcement learning algorithms, such as tabula rasa [28], the target of reaching a goal state has exponential complexity, which is prohabitive in large-scale problems.

Due to the finite-time optimality guarantee and polynomial complexity in implementation, the stochastic MAB as an online learning theory is highly identified for channel resource allocation problems and has been widely applied in dynamic spectrum access [31], [32], decentralized channel access [33], [34] and multi-user *dynamic channel allocation* (DCA) [35]–[37], and so on. Besides, previous works such as [14] have used trajectory prediction and radio maps obtained from real measurement to predict future average data rate. As a result, the scheduling policy is in a single timescale, although the user trajectory, the traffic pattern, and the small timescale CSI are evolving and predictable in different timescales, which naturally requires multiple timescale resource allocations. But none of them have considered the spatial channel distribution variation in wireless communications and 5G networks.

Efforts on the QoE-aware channel allocation scheme based on CSI was seen in [37] for cognitive radio networks and in [22] for video streaming in mobile networks. However, these offline models assume known channel statistics, and do not take into account the user mobility. The motivation of applying contextual MAB for channel allocation in terms of QoE stems from the machine learning society for the applications on ad placements [8], [9] and personalized news recommendation

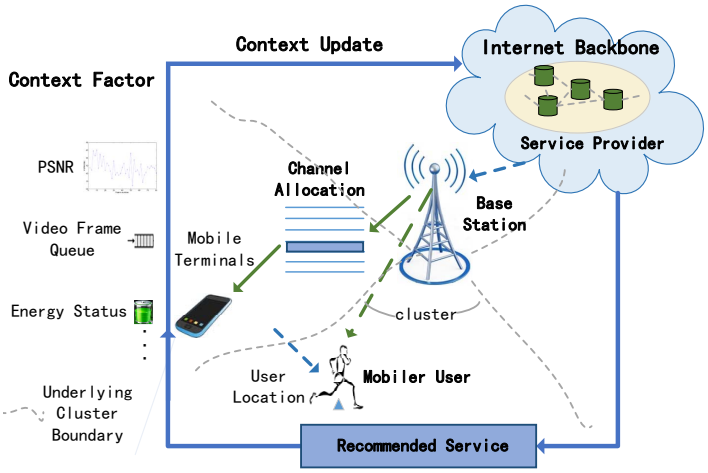


Fig. 1: Human-behavior and QoE-aware Online Learning for DCA in a 5G Downlink

[38] with *arbitrarily* changing contexts and arm distributions. Different to these works, we devise the SCB algorithm based on the well-known UCB1 [11] by considering *stochastically* distributed contexts and channels, which is more suitable for wireless communications. Recently, the work [10] studied a SCB problem in mobile networks with *known* reward function. However, it is not suitable for general QoE-driven applications, where the QoE metric (reward function) is often difficult to model and unavailable *in advance*.

Utilizing human behavioral data in the design loop of mobile networks can improve the network performance. Survey papers like [4] [5] have emphasized its importance and especially human's QoE is the key performance index [6]. In [39], the human behavior is preliminarily involved in video QoE prediction. In [40], the user mobility is predicted based on the human behavior. In [41], a closed-loop framework of data-guided network resource allocation is provided to optimize the users' QoE. In [24], the authors study the mobility and CSI-aware predictive resource allocation for energy-saving based on *future* average user behavioral and channel information. In [4], the authors have pointed out that the massive user behavior data collected via mobile crowdsensing could provide high-throughput, low-delay traffic and energy-efficient services that challenge the limited capacity of wireless networks, but significantly many research problems are still open. Moreover, devising effective multi-user communication protocols for resource sharing is an important issue in 5G networks. Our current work consider both the user location data and QoE-driven contextual data. As illustrated in Fig. 1, for a cellular video streaming service, the *PSNR*, *video frame queueing size* and *energy status* information can be regarded as contexts in the QoE metric for our algorithm to guide the channel resource management. We consider four typical multi-user RA MAC protocols. Therefore, our work can be regarded as a first important attempt for DCA by addressing the challenge that the stochastic channel distributions are location-aware. In addition, we have utilized the technique of grouping the similar user locations, channel conditions and contextual features to further reduce and complexity and speed up the machine

learning (regret) performance.

III. System Model

We consider a typical downlink cellular communication scenario for 5G as shown in Fig. 1. Time is slotted $n = 1, 2, 3, ...t, ...$ and the mobile user's location is unchanged at a time slot but could varies at different time slots. The user transmits its QoE factors as contexts to the base station (BS) at each slot. The BS utilizes the mobile user's QoE and location data, and makes the channel allocation decisions. Note that when the BS makes the channel resource allocation decisions to predict CSI, the potentially incurred delay in reporting the context information to BS in practice will not affect the QoE performance. Because the fast fading CSI is acquired in small timescale when compared to the large timescale less varied context information. The main notations are listed in Table I and Table II.

Denote s as the user's location and $\mathcal{S} = \{1, ..., s, ..., \bar{S}\}$ as the set of all user's collected locations with cardinality $\bar{S}$. Let $\mathcal{N} = \{1, ..., i, ..., N\}$ denote the set of available channels (arms) with cardinality N. Let $\mathcal{X} = \{x_1, x_2..., x_{|\mathcal{X}|}\}$ denote the set of all channels' possible states, and $|\mathcal{X}|$ is the number of channel states. We let $x_{i,t}$ denote the CSI or the value of i-th channel at slot t. Let $\mathbf{y} \in \mathcal{Y}$ is a context vector in the context space $\mathcal{Y}$ that encodes a specific QoE factor. Because we consider the wireless multimedia communications in our experiments, we adopt some of its key context vectors as detailed in Section VII. There can be many other factors that be adopted as the context, e.g., type of applications, type of devices, and type of subscription plans, etc. And, consequently, we can design more user-friendly QoE performance metrics. As we can see, our proposal can be custom-made to capture any different context features.

For clarity, and to highlight the role of latent information, we assume all the three sets are finite, and as such we have discrete[1] context space $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_k, ..., \mathbf{y}_K\}$ with K elements and $\mathcal{K}$ is the index set. Formally, let $\nu = \{\nu_{i,\mathbf{y}_k,s}\}_{i\in\mathcal{N},k\in\mathcal{K},s\in\mathcal{S}}$ be generally the real-valued channel-context-and-location distributions. At each slot $t \in \mathbb{N}$, the user's mobility results in some location $s_t \in \mathcal{S}$ according to some unknown stochastic mobility process $\Upsilon(s_t)$. Then, s_t is revealed to the BS. Observing s_t and the current context vector $\mathbf{y}_k$, an algorithm with policy π must select some $i_t \in \mathcal{N}$ at BS to allocate to the user for transmission at each slot t. Finally, a reward r_t as the realized value of QoE function is sampled from ν and observed for channel i_t with CSI $x_{i,t}$.

In the classic definition of *regret*, the goal is to find up to n slots a sequence of channels $i_{1:n} = \{i_t\}_{1\le t\le n}$ with maximal accumulated expected reward of QoE $\sum_{t=1}^{n} {}_{r_t\sim\nu_{i_t,\mathbf{y}_t,s_t}}[r_t]$.

Let $\mu_{i,k,s} \in \mathbb{R}$ be the mean of $\nu_{i,\mathbf{y}_k,s}$ that is obtained by

$$\mu_{i,k,s} = \mathbb{E}_{r_i\sim\nu_{i,\mathbf{y}_k,s_t}}[r_i] = \mathbb{E}[r(x_i, \mathbf{y}_k; s)], \quad (1)$$

where $r(x_i, \mathbf{y}_k; s)$ denotes the reward under the observed context $\mathbf{y}_k$ at location s, or it is short for $r_{i,k,s}$.

To facilitate online learning and channel resource sharing among multiple users, our goal now is to allocate the channel with the D-th largest expected reward to certain ordered and prioritized users and regard other channels (even channels with larger reward) to be sub-optimal channels, so that the D-th ranked user learns to access the channel with the D-th largest reward. We aim to minimize the number of times that we pick the wrong channel. Here we define two types of regrets as follows:

- *Type* 1 *regret*: The sum of the absolute difference between the expected reward $\mu_{D,\mathbf{y}^D,s}$ that the genie now pick a channel with the D-th largest expected reward and the corresponding optimal context $\mathbf{y}^D$, and the r_t that obtained by the given policy π at each time slot up to slot n, i.e.,
$$\mathcal{R}^1_\pi(s; n) = \sum_{t=1}^{n} |\mu_{D,\mathbf{y}^D,s} - \mathbb{E}_{r_t\sim\nu_{i_t,\mathbf{y}_t,s_t}}[r_t]|. \quad (2)$$
- *Type* 2 *regret*: The absolute difference between the expected sum reward $n\mu_{D,\mathbf{y}^D,s}$ that could be obtained by a genie that pick a channel with D-th largest expected reward and the corresponding optimal context $\mathbf{y}^D$, and that obtained by the given policy π after n plays, i.e.,
$$\mathcal{R}^2_\pi(s; n) = |n\mu_{D,\mathbf{y}^D,s} - \textstyle\sum_{t=1}^{n} \mathbb{E}_{r_t\sim\nu_{i_t,\mathbf{y}_t,s_t}}[r_t]|. \quad (3)$$

Note that $\mathcal{R}^2_\pi(s; n) \le \mathcal{R}^1_\pi(s; n)$ due to the simple fact that $|n\mu_{D,\mathbf{y}^D,s} - \sum_{t=1}^{n} \mathbb{E}_{r_t\sim\nu_{i_t,\mathbf{y}_t,s_t}}[r_t]| \le \sum_{t=1}^{n} |r_{D,\mathbf{y}^D,s} - \mathbb{E}_{r_t\sim\nu_{i_t,\mathbf{y}_t,s_t}}[r_t]|$. Note that, when $D = 1$ which corresponds the typical selection of the *optimal channel*, we have the absolute value sign to be removed and now $\mathcal{R}^1_\pi(s; n) = \mathcal{R}^2_\pi(s; n)$.

The set $\mathcal{S}$ contains user's location points dispersed within the mobile network's region. We model this latent information by assuming that $\mathcal{S}$ is partitioned into C clusters $\mathcal{C} = \{\mathcal{S}_c\}_{c=1,...,C}$ such that the distributions $\nu_c = \{\nu_{i,\mathbf{y}_k,c}\}_{i\in\mathcal{N},k\in\mathcal{K},c\in\mathcal{C}}$ are the same for each $s \in \mathcal{S}_c$. This common distribution ν_c is called a *cluster* distribution. Accordingly, for each $s \in \mathcal{S}_c$, we can defeine $\mathcal{R}^1_\pi(c; n)$ and $\mathcal{R}^2_\pi(c; n)$. In practical mobile networks, both the partitions and the number of clusters are unknown.

Similarly to the UCB1 [11] algorithm, we need to define some important notations for our problem. At time slot n, we denote the number of observations for the triplet $(i, \mathbf{y}_k, s)$ by $T_{i,k,s}(n) = \sum_{t=1}^{n} \mathbb{1}\{i_t = i, \mathbf{y}_t = \mathbf{y}_k, s_t = s\}$, where $\mathbb{1}\{\cdot\}$ is the indicator function. Then, we can define the grouping number of observations $T_{i,s}(n) = \sum_{k\in\mathcal{K}} T_{i,k,s}(n)$ and $T_s(n) = \sum_{i\in\mathcal{N}} T_{i,s}(n)$. We use $\hat{\nu}_{i,\mathbf{y}_k,s}(n)$ and $\hat{r}_{i,k,s}(n)$ to denote the empirical distribution and mean value built from the same observations, respectively. We define R_k, the maximum deviation in the rewards for each context $\forall k \in \mathcal{K}$, as

$$R_k = \sup_{s\in\mathcal{S}}\{\sup_{x\in\mathcal{X}} \mathbb{E}[r(x_i, \mathbf{y}_k; s)] - \inf_{x\in\mathcal{X}} \mathbb{E}[r(x_i, \mathbf{y}_k; s)]\}.$$

We denote the $R = \max_{k\in K} R_k$ as the maximal such deviation among all the contexts. In our problem, we can tailor each UCB1 to a particular pair of user-location and QoE-driven context vector. We denote the $U_{i,k,s}(n)$ a high probability upper confidence bound (UCB) on the mean $\mu_{i,k,s}$, i.e.,

$$U_{i,k,s}(n) = \hat{r}_{i,k,s} + R_k\sqrt{\frac{2\log T_s(n)}{T_{i,s}(n)}}. \quad (4)$$

Correspondingly, we define $L_{i,k,s}(n)$ a high probability lower

[1]This model can be easily extended to continuous parameter settings as [9], [10], [44].

TABLE I: Main Notations

Notations	Definition
t, n	current time slot and total number of slots
$\mathcal{S}, s_t$	set of users' locations and user location at t
$\mathcal{N}$	set of channels with size N with channel index i
$\mathcal{X}$	set of channel states with size $\|\mathcal{X}\|$
$\mathcal{Y}$	set of contexts with size K with context index k
$x_{i,t}$	the CSI or the value of i-th channel at slot t
i_t	the selected i-th channel at t
$\nu = \{\nu_{i,\mathbf{y}_k,s}\}$	channel-context-and-location distributions
r_t	realized reward of QoE at slot t
$r(x_i, \mathbf{y}_k; s)$	QoE with CSI x_i under context $\mathbf{y}_k$ at location s
$\mu_{i,k,s}$	mean of $\nu_{i,\mathbf{y}_k,s}$
$\mathcal{R}^1_\pi(s;n), \mathcal{R}^2_\pi(s;n)$	Type 1 and tyep 2 regret for location s
$\mathcal{R}^1_\pi(c;n), \mathcal{R}^2_\pi(c;n)$	Type 1 and tyep 2 regret for cluster c
$C, \mathcal{S}_c$	set of clusters and the type c cluster
$T_{i,k,s}(n)$	number of observed triplet $(i, \mathbf{y}_k, s)$
$T_{i,s}(n)$	number of observed $(i, \mathbf{y}_k, s)$ over all $\mathbf{y}_k$
$T_s(n)$	number of observed $(i, \mathbf{y}_k, s)$ over all i and $\mathbf{y}_k$
$\hat{\nu}_{i,\mathbf{y}_k,s}(n), \hat{r}_{i,k,s}(n)$	the empirical distribution and mean value
R_k	maximum deviation in the rewards for context k
$U_{i,k,s}(n), L_{i,k,s}(n)$	UCB and LCB for triplet $(i, \mathbf{y}_k, s)$
$H_{i,k,s}(n), G_{i,k,s}(n)$	HPCI and confidence gap for triplet $(i, \mathbf{y}_k, s)$
$U_{i,k,c}(n), L_{i,k,c}(n)$	UCB and LCB for triplet $(i, \mathbf{y}_k, c)$
$T_{i,k,c}(n)$	number of observed triplet $(i, \mathbf{y}_k, c)$
$\hat{r}_{i,k,c}(n)$	empirical mean for triplet $(i, \mathbf{y}_k, c)$
$T_{i,c}(n)$	number of observed $(i, \mathbf{y}_k, c)$ over all $\mathbf{y}_k$
$T_c(n)$	number of observed $(i, \mathbf{y}_k, c)$ over all i and $\mathbf{y}_k$
$\mathcal{A}_{D,c}$	channel set with D-th largest reward in cluster c
$T_{D,i,c}(n)$	number of times BS allocates $i \notin \mathcal{A}_{D,c}$ up to n
$\mu_{D,k,c}$	D-th largest expected reward under $\mathbf{y}_k$ in c
$\mu_{i,k,c}$	i-th expected reward for $i \notin \mathcal{A}_{D,c}$ under $\mathbf{y}_k$ in c
$\Delta^k_{D,i,c}$	absolute gap between $\mu_{D,k,c}$ and $\mu_{i,k,c}$
$\Delta_{\max,c}$	$\max_{i,k} \Delta^k_{D,i,c}$
$r_{D,\mathbf{y}(t),c}(n)$	reward for selected triplet $(D, \mathbf{y}_k, c)$ at slot t
$\hat{\Delta}^k_{D_c,i,c}$	pseudo optimality gaps in (12)
$\epsilon = \epsilon_{i,k,s,s',n}$	adaptive factor for enlarged confidence bounds
$\mathfrak{S}_s(n)$	compatible set or *clique* of user location graph
$\mathfrak{S}^+_s(n)$	maximally compatible set or maximal cliques
$\Omega_n, \mathcal{F}^\alpha_n, \mathcal{E}^\alpha_n$	$\mathcal{S}_{c_n} \in \mathfrak{S}_{s_n}(n-1)$ and events in (13) and (14)
$T_{D_{c_n},i_n,s_n}(n-1)$	number of selecting $i \notin \mathcal{A}_{D,c}$ at s_n up to $n-1$
$T_{D_{c_n},i_n,\mathcal{S}_{c_n}}(n-1)$	number of selecting $i \notin \mathcal{A}_{D,c}$ at $\mathcal{S}_{c_n}$ up to $n-1$
γ_c	distortion factor
$\mathcal{S}_c(s;\gamma)$	γ-balance of $\mathcal{S}$
$\Upsilon(s), \Upsilon$	user arrival distribution and uniform distribution
$\mathcal{R}^{1,2}_{SCB(D)\ on\ S}(n)$	Type 1 and tyep 2 regret for $SCB(D)$ on S
$\mathcal{R}^{1,2}_{SCB(D)\ on\ C}(n)$	Type 1 and tyep 2 regret for $SCB(D)$ on C
$\mathcal{M}$	set of user with size M with user index j or m
$\mathcal{R}_\pi(n)$	regret for multiple users
$\mathcal{O}^*_M$	set of M channels with M largest expected rewards
$\mathcal{O}^*_m$	set of channels with m-th largest expected rewards
$\hat{r}_{j_t,k,T_{j_t,k,c}(t-1),c}$	empirical mean under $(j, \mathbf{y}_k, c)$ up to $T_{j_t,k,c}(t-1)$

confidence bound (LCB), i.e.,

$$L_{i,k,s}(n) = \hat{r}_{i,k,s} - R_k \sqrt{\frac{2\log T_s(n)}{T_{i,s}(n)}}. \tag{5}$$

Let the high probability confidence interval (HPCI) be $H_{i,k,s}(n) = [L_{i,k,s}(n), U_{i,k,s}(n)]$ and the confidence gap be $G_{i,k,s}(n) = U_{i,k,s}(n) - L_{i,k,s}(n)$.

IV. The SCB Problem with Single User's Known Mobility within a Cluster

In this section, we consider the ideal case that the mobile user's arrivals, i.e., locations $\{s_n\}_{n\geq 1}$, up to current slot n belong to a same *known* cluster $c \in \mathcal{C}$. The different locations of the user within cluster c share the same channel-and-context distribution $\{\nu_{i,\mathbf{y}_k,c}\}_{i\in\mathcal{N},k\in\mathcal{K}}$. In this case, the uncertain latent information of the underlying cluster is removed. Thus, the dynamic channel allocation problem degenerates to a SCB problem that needs to keep only a single learning instance for the cluster. Indeed, if two distributions $\{\nu_{i,\mathbf{y}_k,s}\}$ and $\{\nu_{i,\mathbf{y}_k,s'}\}$ are the same, then group the corresponding observations provides a faster convergence speed. This setting provides great insight for our rest discussions.

A. The SCB(D) algorithm

We first propose a general policy to allocate a channel with the D-largest expected reward ($1 \leq D \leq N$) for the single cluster SCB problem. It is summarized as SCB(D) in Alg. 1. The motivation is that this policy can facilitate decentralized sharing of channels among multiple users in Section VI, such that user m will run a learning policy targeting a channel with the m-th largest expected reward.

The SCB(D) generalizes the UCB1 algorithm [11] in two aspects: First, the BS keeps a UCB1 instance for each observed context vector variable $\mathbf{y}_k$ and select the channel $i \in \mathcal{N}$ that is the minimal one in $L_{i,k,c}(n)$ with the set $\mathcal{O}_{D,c}$ containing D channels with the D-largest empirical channel reward that maximizes $U_{i,k,c}(n)$ at slot n. Both $U_{i,k,c}(n)$ and $L_{i,k,c}(n)$ have similar definitions as in (4) and (5) for the cluster c, where $T_{i,k,c}(n) = \sum_{t=1}^n \mathbb{1}\{i_t = i, \mathbf{y}_t = \mathbf{y}_k, s_t \in \mathcal{S}_c\}$ and the corresponding $T_{i,c}(n)$ and $T_c(n)$.

The key idea is to store an empirical mean of $\mu_{i,k,c}$ for every channel-context pair for the cluster c, i.e., $\hat{r}_{i,k,c}$. Specifically, the BS needs to use an $N \times K$ matrix $[\hat{r}_{i,k,c}]_{N\times K}$ to store the accumulated empirical reward from previous channel allocation decisions. The BS also needs to store $[T_{i,c}(n)]_{1\times N}$, $\forall i \in \mathcal{N}$ up to the current slot n. At the slot n, the channel with index i_n is allocated under the observed context $\mathbf{y}_k$ and reward $r(x_{i_n}, \mathbf{y}_k; c)$ is revealed. Based on these information, $[\hat{r}_{i,k,c}]_{N\times K}$ and $[T_{i,c}(n)]_{1\times N}$ are updated. Since the BS always keeps the accumulated values, the storage complexity is only $\Theta(NK)$ that does not increase with time. And, obviously, the overall computational complexity is $O(NKn)$.

B. Regret Analysis for SCB(D)

In classic MAB-based channel allocation problems [33], [34], [36], [37], the regret is calculated based on the sub-optimal allocated channels and it is upper bounded by the summation of optimality gaps times the expected number of pulls for each non-optimal channels. Now, we present the analysis of the regret upper bound for SCB(D). Denote $\mathcal{A}_{D,c}$ as the set of channels with the D-th largest expected reward in the cluster c. We allow multiple channels with the D-th largest expected reward, and all these channels are regarded as optimal channels. Moreover, an optimal channel could correspond to several context values.

Lemma 1. Under the policy SCB(D), the expected number of the times that the BS allocates $i \notin \mathcal{A}_{D,c}$ up to time n has:

$$\mathbb{E}[T_{D,i,c}(n)] \leq \frac{8R^2 \ln n}{\min\limits_{1\leq k\leq K} \left(\Delta^k_{D,i,c}\right)^2} + 1 + \frac{2\pi^2}{3}, \tag{6}$$

where $\Delta^k_{D,i,c} = |\mu_{D,k,c} - \mu_{i,k,c}|$ and $\mu_{D,k,c}$ and $\mu_{i,k,c}$ are the D-th largest expected reward and the i-th expected reward with $i \notin \mathcal{A}_{D,c}$ under context $\mathbf{y}_k$ in cluster c, respectively.

Proof: See detailed proof in [48]. ■

Algorithm 1: SCB(D): Channel Allocation for the D-Largest Expected Reward

Input: Receive $s_n \sim \Upsilon$. Map the mobile user locations $\{s_n\}_{n\geq 1}$ to a cluster $\mathcal{S}_c$
Initialization $[\hat{r}_{i,k,c}]_{N\times K} = [\mathbf{0}]_{N\times K}$ and $[T_{i,c}(n)]_{1\times N} = [\mathbf{0}]_{1\times N}$
The BS observes the current context $\mathbf{y}_n = \mathbf{y}_k$.
for $n = 1$ to N **do**
  Let $i = n$ and select channel $i_n = i$;
  Update $\hat{r}_{i,k,c} = r(x_{i_n}, \mathbf{y}_k; c), \forall k : 1 \leq k \leq K$;
  $T_{i,c}(n) = 1$
end for
while 1 **do**
  $n = n + 1$
  Let $\mathcal{O}_{D,c}$ contains the D channels with the D largest values in $U_{i,k,c}(n), \forall i \in \mathcal{N}, \forall k \in \mathcal{K}$
  Allocate channel d in $\mathcal{O}_{D,c}$, such that,
$$d = arg\min_{i\in\mathcal{O}_{D,c}} L_{i,k,c}(n).$$
  Update $\hat{r}_{i,k,c}$ and $T_{i,c}(n), \forall 1 \leq k \leq K, \forall s_n \in \mathcal{S}_c$ as
$$\hat{r}_{i,k,c}(n) = \frac{\hat{r}_{i,k,c}(n-1)T_{i,c}(n-1)+r(x_{i_n},\mathbf{y}_k;s_n)}{T_{i,c}(n-1)+1},$$
$$T_{i,c}(n) = T_{i,c}(n-1) + 1$$
end while

Based on Lemma 1 and using the simple fact that $\mathcal{R}^2_\pi(c;n) \leq \mathcal{R}^1_\pi(c;n)$, we can easily provide the regret bound in the following theorem:

Theorem 1. The expected regret $\mathcal{R}_\pi(c;n)$ under both regret definitions in (2) and (3) for cluster c for SCB(D) is upper bounded by

$$\sum_{i:i\notin\mathcal{A}_{D,c}} \frac{8R^2\Delta_{\max,c}\ln n}{\min\limits_{1\leq k\leq K}\left(\Delta^k_{D,i,c}\right)^2} + \sum_{i:i\notin\mathcal{A}_{D,c}} \left(1+\frac{2\pi^2}{3}\right)\Delta_{\max,c}, \tag{7}$$

where $\Delta_{\max,c} = \max\limits_{i,k} \Delta^k_{D,i,c}$.

Proof: Under the policy SCB(D), we have:

$$\begin{aligned}\mathcal{R}^2_\pi(c;n) &\leq \mathcal{R}^1_\pi(c;n)\\ &= \textstyle\sum_{t=1}^n |r_{D,\mathbf{y}(t),c} - \mathbb{E}_{r_s\sim\nu_{i_t,\mathbf{y}_t,s_t}}[r_t]|\\ &\leq \sum_{i:i\notin\mathcal{A}_{D,c}} \Delta_{\max,c}\mathbb{E}[T_{D,i,c}(n)]\\ &\leq \sum_{i:i\notin\mathcal{A}_{D,c}} \frac{8R^2\Delta_{\max,c}\ln n}{\min_{1\leq k\leq K}\left(\Delta^k_{D,i,c}\right)^2} + \sum_{i:i\notin\mathcal{A}_{D,c}}\left(1+\frac{2\pi^2}{3}\right)\Delta_{\max,c}.\end{aligned}$$

■

Since each channel's contribution to the regret is logarithmic in time and they are explored independently for all contexts, the regret upper bound scales as $O(NKlog(n))$.

V. The LCMAB Problem with Single User's Unknown Mobility

In this section, we come to the general case that the user's mobility is not restricted and can vary across *multiple* different underlying clusters, and the boundaries of clusters and their distributions are not known to the BS. We first indicate that it is a very challenging problem according to the following high level intuition.

Let us focus on the means of reward only. Let $\mathcal{C}_{n-1} = \{c \in \mathcal{C}, \forall i \in \mathcal{N}, \forall \mathbf{y}_k \in \mathcal{Y} : \mu_{i,k,c} \in H_{i,k,\mathcal{S}}(n-1)\}$ be the set of admissible clusters viewed as BS at slot $n-1$, where the confidence set $H_{i,k,\mathcal{S}}(n-1)$ is constructed using observations for the triplet $\{(i,\mathbf{y}_k,b)\}_{b\in\mathcal{S}}$. Note that the true cluster c that the user at current location belonging to is admissible due to the concentration of measure, and thus $\mathcal{C}_{t-1}$ is not empty. Let $\tilde{c} \in \mathcal{C}_{t-1}$ be an admission cluster. Then, the SCB(D) in Alg. 1 would allocate a channel $D_{\tilde{c}}$ with D-th largest expected reward $\mu_{D,k,\tilde{c}}$, where the simple notation k denotes the optimal context such that its value given the least expected regret under SCB(D). Now there are several possible situations:

1) : If another cluster $c' \in \mathcal{C}$ such that $\mu_{D_{\tilde{c}},k,c'} - \mu_{D_{\tilde{c}},k,\tilde{c}} > H_{i,k,\mathcal{S}}(t-1)$, the cluster c' cannot be admissible. Because $\mu_{D,k,c'}$ is out of the possible confidence range under SCB(D). The c' is admissible only if $D_{\tilde{c}} = D_{c'}$. It means that choosing to play $D_{\tilde{c}}$ for $c' \in \mathcal{C}_{n-1}$ does no cause harm.

2) : If $\exists c' \in \mathcal{C}$ such that both $\mu_{D_{\tilde{c}},k,c'} - \mu_{D_{\tilde{c}},k,\tilde{c}} \leq H_{i,k,\mathcal{S}}(t-1)$ and $D_{\tilde{c}} \neq D_{c'}$, standing for the case that the channel state distribution of the channel set and the latent distribution of different clusters c' and $\tilde{c}$ are very close to each other, then the implemented SCB(D) on BS cannot discriminate the right and wrong channels very well. In practice, this corresponds to the wireless environment that has very good channel condition, and the users' mobility are very restricted and almost static. Here, we note that if we choose the D-th largest reward in cluster $\tilde{c}$, it guarantees that $|\mu_{D_c,k,c} - \mu_{D_{\tilde{c}},k,\tilde{c}}| \leq |\mu_{D_{\tilde{c}},k,\tilde{c}} - \mu_{D_{\tilde{c}},k,c}|$, which leads to a *controlled error*. This is due to the difference of channel and context profiles are small between $\tilde{c}$ and c, where this phenomenon is verified in our real experiments.

Under known cluster information, we can obtain significantly improved regret bounds on the equivalent agnostic version. Indeed, if two distributions $\nu_{i,\mathbf{y}_k,s}$ and $\nu_{i,\mathbf{y}_k,s'}$ are the same, then grouping the corresponding observations provides a faster convergence speed. Now, we try to resolve the much harder agnostic version that cluster distributions are unknown with significantly improved regret bounds.

A. Grouping Channel-Context-and-Location Distributions

Before presenting the agnostic version of SCB(D) algorithm to adaptively learn the underlying clusters at the BS, we first discuss grouping channel distributions for speeding up the learning. By our analysis, the grouping problem is modelled as *maximal clique* in a graph covering problem.

For a subset of locations $S \subset \mathcal{S}$, define the empirical group distribution $\hat{\nu}_{i,k,S}(n)$ with associated group mean $\mu_{i,k,S}(n)$, confidence intervals $U_{i,k,S}(n)$, $L_{i,k,s}(n)$ and set $H_{i,k,S}(n)$, where

$$\hat{\nu}_{i,k,S}(n) = \frac{\sum_{s'\in\mathcal{S}}\hat{\nu}_{i,k,s'}(n)T_{i,k,s'}(n)\mathbb{1}\{s'\in S\}}{\sum_{s'\in\mathcal{S}}T_{i,k,s'}(n)\mathbb{1}\{s'\in S\}},$$
$$\mu_{i,k,S}(n) = \frac{\sum_{s'\in\mathcal{S}}\mu_{i,k,s'}(n)T_{i,k,s'}(n)\mathbb{1}\{s'\in S\}}{\sum_{s'\in\mathcal{S}}T_{i,k,s'}(n)\mathbb{1}\{s'\in S\}}.$$

When $S = \mathcal{S}_c$, then $\mu_{i,k,\mathcal{S}_c}(n) = \mu_{i,k,c}$, which may not hold for other sets $\hat{S}$ because a bias occurs when the $\{\mu_{i,k,s'}\}_{s'\in\hat{S}}$ are distinct. Further, the convergence speed of the group depends on $T_{i,k,S}(n) = \sum_{s'\in\mathcal{S}} T_{i,k,s'}(n)\mathbb{1}\{s'\in\mathcal{S}\}$, which is typically much faster than that of a single location s that depends on $T_{i,k,s}(n)$. Thus, $H_{i,k,S}(n) = [L_{i,k,S}(n), U_{i,k,S}(n)]$

is potentially much smaller than $H_{i,k,s}(n)$. Another important fact is, $\mu_{i,k,S}(n) \in H_{i,k,S}(n)$ holds with high probability, but for some $s \in S$ there is no reason that $\mu_{i,k,s} \in H_{i,k,S}(n)$ due to the introduced bias.

To leverage the estimation bias, we restrict possible groups S by using two observations. First, if $\mu_{i,k,s} = \mu_{i,k,s'}$, then we must have $H_{i,k,s}(n) \cap H_{i,k,s'}(n) \neq \emptyset$ with high probability. More generally, it there is some S such that $\mu_{i,k,s} = \mu_{i,k,s'}$, for all $s, s' \in S$, it must satisfy that for all $S' \subset S$ and all $S'' \subset S$, with high probability, $H_{i,k,S''}(n) \cap H_{i,k,S'}(n) \neq \emptyset$. Second, we define an adaptive factor $\epsilon = \epsilon_{i,k,s,s',n}$ for the enlarged confidence bounds

$$\begin{aligned} U_{i,k,s}(n;1+\epsilon) &= \hat{r}_{i,k,s}(n) + (1+\epsilon)(U_{i,k,s}(n) - \hat{r}_{i,k,s}(n)), \\ L_{i,k,s}(n;1+\epsilon) &= \hat{r}_{i,k,s}(n) + (1+\epsilon)(L_{i,k,s}(n) - \hat{r}_{i,k,s}(n)). \end{aligned} \tag{8}$$

Consequently, we have $H_{i,k,s}(n;1+\epsilon) = [L_{i,k,s}(n;1+\epsilon), U_{i,k,s}(n;1+\epsilon)]$. If $\mu_{i,k,s} = \mu_{i,k,s'}$, we have $G_{i,k,s'}(n) \leq \epsilon \min\{U_{i,k,s}(n) - \hat{r}_{i,k,s}(n), L_{i,k,s}(n) - \hat{r}_{i,k,s}(n)\} = \frac{\epsilon}{2} G_{i,k,s}(n)$. Then, we must have $H_{i,k,s'}(n) \subset H_{i,k,s}(n, 1+\epsilon)$ with high probability.[2]

The distribution $\nu_{i,\mathbf{y}_k,s}$ for each context vector $\mathbf{y}_k$ never changes at different channel $i \in \mathcal{N}$ and user locations $s \in \mathcal{S}$. Thus, grouping all the context vectors does not affect the differentiation of the underlying clusters and can further speed up the performance during the process of clustering. Therefore, we introduce the following grouped mean $\mu_{i,S}(n)$, grouped estimated mean values $\hat{r}_{i,S}(n)$.

$$\mu_{i,S}(n) = \frac{\sum_{\mathbf{y}_{T_{i,k,s'}(n)} = \mathbf{y}_k \in \mathcal{Y}} \sum_{s' \in \mathcal{S}} \mu_{i,k,s'}(n) T_{i,k,s'}(n) \mathbb{1}\{s' \in S\}}{\sum_{\mathbf{y}_{T_{i,k,s'}(n)} = \mathbf{y}_k \in \mathcal{Y}} \sum_{s' \in \mathcal{S}} T_{i,k,s'}(n) \mathbb{1}\{s' \in S\}},$$

$$\hat{r}_{i,S}(n) = \frac{\sum_{\mathbf{y}_{T_{i,k,s'}(n)} = \mathbf{y}_k \in \mathcal{Y}} \sum_{s' \in \mathcal{S}} r(x_i, \mathbf{y}_k; s'_n)(n) T_{i,k,s'}(n) \mathbb{1}\{s' \in S\}}{\sum_{\mathbf{y}_{T_{i,k,s'}(n)} = \mathbf{y}_k \in \mathcal{Y}} \sum_{s' \in \mathcal{S}} T_{i,k,s'}(n) \mathbb{1}\{s' \in S\}}.$$

Similarly, we have the grouped lower confidence bound $L_{i,S}(n)$ and upper confidence bound $U_{i,S}(n)$, and the confidence interval $H_{i,S}(n) = [L_{i,S}(n), U_{i,S}(n)]$ is probably much smaller than $H_{i,k,S}(n)$. Due to the linear additive property of all context variables defined in (4) and (5) such that in $L_{i,S}(n) = \sum_{\mathbf{y}_k \in \mathcal{Y}} L_{i,k,S}(n)$ and $U_{i,S}(n) = \sum_{\mathbf{y}_k \in \mathcal{Y}} U_{i,k,S}(n)$, we still have $\mu_{i,S}(n) \in H_{i,S}(n)$ with high probability.

Define the *compatible set* as

$$\begin{aligned} \mathfrak{S}_s(n) \triangleq \Big\{ & S \subset \mathcal{S} : \forall i \in \mathcal{N}, \\ & \forall s', s'' \in S, H_{i,s'}(n) \subset H_{i,s''}(n;1+\epsilon) \\ & \forall S', S'' \subset S, H_{i,S''}(n) \cap H_{i,S'}(n) \neq \emptyset \Big\}, \end{aligned}$$

and the *maximally compatible set* that has the maximal group speed of convergence and a controlled bias as

$$\mathfrak{S}_s^+(n) \triangleq \text{Argmax}_{S \in \mathfrak{S}_s(n)} S. \tag{11}$$

Note that contrary to *argmax*, here *Argmax* returns a set rather than a real value.

The set $\mathfrak{S}_s(n)$ can be viewed as a *clique* in the graph theory that covers the node s by viewing the user locations as nodes in the graph $\mathfrak{S}(n)$. The problem of determining the maximally compatible sets $\mathfrak{S}_s^+(n)$ as *maximal cliques* is converted as a graph covering problem. BS can use existing greedy set cover algorithm (e.g., [46], p. 16) to solve the above problem.

[2]In this paper, for clarity, we only focus on mean-based confidence bound analysis for our stochastic MAB problem, although other approaches like to use the empirical distributions $\hat{\nu}_{i,k,s}(n)$ to measure s' with obvious mismatch in Kullback-Leibler divergence are possible.

Algorithm 2: A-SCB(D): Channel Allocation with Unknown Cluster Distributions

1: **Require:** Parameter γ.
2: **for** $n = 1...$ **do**
3: Receive the user's location $s_n \sim \Upsilon$ and observe the context $\mathbf{y}_k$ at BS
4: Compute $\hat{r}_{i,k,s}(n-1)$, then $U_{i,k,s}(n-1)$, $L_{i,k,s}(n-1)$, $H_{i,k,s}(n-1)$ and $H_{i,k,s}(n-1)$.
5: Define the quantity $\epsilon = \epsilon_{s_n,s',n-1}$ by
$$\max\left\{\sqrt{\frac{2\gamma \log(T_{s'}(n-1))}{\log(T_{s_n}(n-1))}} - 1, 0\right\}. \tag{9}$$
6: Compute $\hat{r}_{i,s}(n-1)$, then $U_{i,s}(n-1)$, $L_{i,s}(n-1)$, $H_{i,s}(n-1)$ and $H_{i,s}(n-1)$, based on the maximally compatible aggregation sets.
7: Let $\mathcal{O}_{\mathfrak{S}_{s_n}^+}$ contains the D channels with the D largest values, given $\mathbf{y}_n = \mathbf{y}_k$ such that for $i \in \mathcal{O}_{\mathfrak{S}_{s_n}^+}$
$$\max_{S \in \mathfrak{S}_{s_n}^+(n-1)} U_{i,k,S}(n-1). \tag{10}$$
8: Allocate channel d in $\mathcal{O}_{\mathfrak{S}_{s_n}^+}$, such that,
$$d = arg \min_{i \in \mathcal{O}_{\mathfrak{S}_{s_n}^+}} \min_{S \in \mathfrak{S}_{s_n}^+(n-1)} L_{i,k,S}(n-1).$$
9: Update $\hat{r}_{i,k,S}$ and $T_{i,S}(n), \forall 1 \leq k \leq K, \forall s_n \in \mathcal{S}$.
10: **end for**

B. The Agnostic SCB(D) Algorithm

We are now ready to introduce A-SCB(D), whose pseudocode is provided as Alg. 2. Proving strong regret bounds in this agnostic setting is difficult without further assumptions on the communication environment, since the cluster size and the admissible true cluster may change at each single time slot. For this reason, we need to make reasonable assumptions to get strong and clear regret bounds in practice.

At first, we obtain Proposition 1 that controls the number of pulls of sub-optimal channels under some potentially high probability events.

Proposition 1. Define the event $\Omega_n = \left\{\mathcal{S}_{c_n} \in \mathfrak{S}_{s_n}(n-1)\right\}$ that the true class c_n is admissible at slot n, where $c_n = c$ is the current cluster that the user located in and $i_n = i$ is the chosen channel. Under Ω_n, there exists a maximally compatible supper set $\tilde{\mathcal{S}}_{c_n}$ that contains $\mathcal{S}_{c_n}$. Now let us define the pseudo optimality gaps under each context $k \in \mathcal{K}$

$$\begin{aligned} \widetilde{\Delta}_{D_c,i,c}^k = \max\Big\{ & \mu_{i,k,c(s)} - \mu_{i,k,c} : s \in \tilde{S}_c \\ & \cap \max_{S \in S_{s_n}(n-1)} U_{i,k,S} \geq U_{D,k,S}, \\ & \mu_{i,k,c} - \mu_{i,k,c(s)} : s \in \tilde{S}_c \\ & \cap \min_{S \in S_{s_n}(n-1)} L_{i,k,S} \leq L_{D,k,S} \Big\}. \end{aligned} \tag{12}$$

Define $\mathcal{F}_n^\alpha$ and $\mathcal{E}_n^\alpha$ as the event that the confidence interval of the optimal channel of cluster $\mathcal{S}_c$ is small enough (that ensures

the $\tilde{S}_c$ is not too biased to the true cluster),

$$\mathcal{F}_n^\alpha = \left\{\alpha\widetilde{\Delta}^k_{D_c,i,c} \le G_{i,k,\mathcal{S}_c}, k \in \mathcal{K}\right\} \quad (13)$$

$$\mathcal{E}_n^\alpha = \{\min\{G_{D_c,k,S_c}(n-1), \max_{s\in\tilde{S}} |\mu_{D,k,c(s)} - \mu_{D,k,c}|\} < \alpha\widetilde{\Delta}^k_{D_c,i_n,c}\}. \quad (14)$$

The events $\mathcal{F}_n^\alpha$ and $\mathcal{E}_n^\alpha$ hold for small $\alpha \in (0,1)$. Under $\Omega_n \cap \mathcal{E}_n^\alpha$ and $\eta \in (\alpha, 1]$, at least one of the following two inequalities are satisfied.

$$T_{D_{c_n},i_n,s_n}(n-1) < \frac{(1+\frac{\epsilon}{2})^2 8R^2\ln(T_{s_n}(n-1))}{(\eta-\alpha)^2 \min\limits_{1\le k\le K}\left(\Delta^{k_n}_{D_c,i_n,c_n}\right)^2} + O(1), \quad (15)$$

$$T_{D_{c_n},i_n,\mathcal{S}_{c_n}}(n-1) < \frac{8R^2\ln(T_{\mathcal{S}_{c_n}}(n-1))}{(1-\eta)^2 \min\limits_{1\le k\le K}\left(\Delta^{k_n}_{D_c,i_n,c_n}\right)^2} + O(1). \quad (16)$$

Proof: See detailed proof in [48]. ∎

Remark 1. Proposition 1 shows that in all the events this occurs with high probabilities (proved in Lemma 2 and 3), the total number of allocated suboptimal channels for either the current user location s_n or its cluster c_n is controlled in a logarithmic order. In particular, for small ϵ, α and $\eta \to 1$, it shows that under $\Omega_n \cap \mathcal{E}_n^\alpha$ the regret of A-SCB(D) is essentially in between that of SCB(D) on $\mathcal{S}$ and SCB(D) on $\mathcal{C}$ up to constants, it is never worse than SCB(D) on $\mathcal{S}$, and can be significantly better by competing occasionally with SCB(D) on $\mathcal{C}$. It now remains to show that $\Omega_n \cap \mathcal{E}_n^\alpha$ happens *with high probability* to deduce a non-trivial regret bound.

1) Adaptive enlargement of ϵ: We introduce an adaptive ϵ and not just a constant $\epsilon = 1$, since a constant ϵ does not always ensure that $\mathcal{S}_{c_n}$ is admissible such that Ω_n happens with high probability, but only that a subset of $\mathcal{S}_{c_n}$ is admissible at slot n. To better understand the number of such locations that are gathered in $H_{i,k,s}(n; 1+\epsilon)$, we define the *distortion factor* which only depends on the law of user arrivals Υ:

Definition 1. The *distortion factor* of group $\mathcal{S}_c$ is defined as

$$\gamma_c = \frac{\max_{s\in\mathcal{S}_c}\Upsilon(s)}{\min_{s\in\mathcal{S}_c}\Upsilon(s)},$$

where $\mathcal{S}_c(s;\gamma) = (s' \in \mathcal{S}_c : \Upsilon(s) \le \gamma\Upsilon(s'))$ is the γ-balance of $\mathcal{S}$ with respect to cluster c for point $s \in \mathcal{S}_c$. ∎

The factor enables us to quantify the effective number of user locations that are grouped with $s \in \mathcal{S}$, which directly indicates the speed-up that the algorithm can achieve. Importantly, if $\gamma \ge \gamma_c$, then it holds that $\mathcal{S}_c(s;\gamma) = \mathcal{S}_c$ for all $s \in \mathcal{S}_c$. A-SCB(D) uses an adaptive ϵ that ensures that if γ is essentially greater that γ_c, then $\mathcal{S}_c(s;\gamma)$ and thus $\mathcal{S}_C$ are admissible with high probability.

2) Ensuring Events Ω_n, $\mathcal{F}_n^\alpha$ and $\mathcal{E}_n^\alpha$ with High Probability: For event Ω_n that the true cluster is admissible, we have the following lemma according to [12] to show that Ω_n is a high probability event.

Lemma 2 [12]. In A-SCB(D), let $\epsilon = \epsilon_{s_n,s',n-1} = \max\left\{\sqrt{\frac{2\gamma\log(T_{s'}(n-1))}{\log(T_{s_n}(n-1))}} - 1, 0\right\}$, then it holds that

$$\mathbf{Pr}(\Omega_n) \le 1 - O\Big(n^{-2}N\sum_{s\in\mathcal{S}}\Upsilon(s)^{-2}\Big) - 2|\mathcal{S}|n^{-2},$$

if $\gamma \ge \gamma_c + O(n^{-1/2})$. ∎

For event $\mathcal{F}_n^\alpha$ that the allocation of suboptimal channels and the errors of miss identifying clusters are controlled and event $\mathcal{E}_n^\alpha$ that the optimal channel is allocated with enough time, we obtain the following lemma to show that under mild conditions and restrictions in mobile networks, they actually hold with high probability.

Lemma 3. Assuming that Υ is the uniform distribution over the network region, and all clusters have the same size. If A-SCB(D) run with $\gamma \sim \gamma_c = 1$, then a) $\mathbf{Pr}(\mathcal{F}_n^\alpha) \ge 1 - O(n^{-2})$ holds for $\alpha \in [1/2, 1]$; b) $\mathbf{Pr}(\mathcal{E}_n^\alpha) \ge 1 - O(n^{-2})$ holds for $\alpha = 1/2$ and that $\forall c, c' \in \mathcal{C}$, $\forall i \in \mathcal{N}, \forall k \in \mathcal{K}$ $\mu_{D_c,k,c} - \mu_{D_c,k,c'} < \Delta^k_{D_c,i,c}/2$ or $\mu_{D_c,k,c} - \mu_{D_c,k,c'} > \frac{3}{2}\Delta^k_{D_c,i,c}$. In other words, a misidentification between two clusters is either clear or harmless.

Proof: See detailed proof in [48]. ∎

Note that obtaining strong regret results in general network conditions is a hard issue. Now, summarizing all the above analysis, we can get the following Theorem 2 for a specific but typical network scenario.

Theorem 2. Assuming that Υ is the uniform distribution, and all clusters have the same size, in this case, if A-SCB(D) run with $\gamma \sim \gamma_c = 1$ and for some $\alpha \in (1/2, 1]$ and $\eta \in (\alpha, 1]$ with ϵ defined in (9), the expected regret defined in (2) and (3) at slot n of the A-SCB(D) satisfies at least one of the following two inequalities:

$$\mathcal{R}^{1,2}_{SCB(D)\,on\,S}(n) \le \sum_{s\in\mathcal{S}}\sum_{i:i\notin A_{D_c,s}} \frac{(1+\frac{\epsilon}{2})^2 8R^2\Delta_{\max,s}\ln(n\Upsilon(s))}{(\eta-\alpha)^2 \min\limits_{1\le k\le K}\left(\Delta^k_{D_c,i,c}\right)^2} + O(1),$$

$$\mathcal{R}^{1,2}_{SCB(D)\,on\,C}(n) \le \max\Bigg\{\sum_{c=1}^{C}\sum_{i:i\notin A_{D,c}} \frac{8R^2\Delta_{\max,c}\ln(n\Upsilon(\mathcal{S}_c))}{(1-\eta)^2 \min\limits_{1\le k\le K}\left(\Delta^k_{D,i,c}\right)^2} + O(1), \sum_{c=1}^{C}\sum_{i:i\notin A_{D,c}} \frac{8R^2\Delta_{\max,c}\ln(n\Upsilon(\mathcal{S}_c))}{\alpha^2 \min\limits_{1\le k\le K}\left(\tilde{\Delta}^k_{D,i,c}\right)^2} + O(1)\Bigg\}.$$

From the regret results of Theorem 2, we find that the regret upper bounds of SCB(D) on S and SCB(D) on C are still in the optimal order of $O(log(n))$. But SCB(D) on C has obvious smaller regret result than SCB(D) on S due to grouping data of same distributions to speeding the online learning process.

VI. Distributed Learning among Multiple Users

In practical wireless networks, there multiple users access the channel resources randomly at the same time by sharing the set of N channels. The multi-access control protocol of users' RA scheme can be both distributed (e.g. CSMA, ALOHA) or centralized (e.g., communicating with BS, TDMA, SIC).

In this section, we consider $\mathcal{M} = \{1, ...j, ..., M\}$ users which may have different distributions $\nu_{i,\mathbf{y}_k,s,j}$, $\forall j \in \mathcal{M}$ due to different user-specific context distributions for different applications.

For channel i allocated to user j at slot t, the QoE reward $r_t = r_{i,k,s}$ under the observed context $\mathbf{y}_k$ at location s is obtained if there is not any other user allocated the same channel. If there are multiple users accessing the same channel, there are three cases:

1) **Contention model** ($\mathbf{M}_1$): At most one of the conflicting users j' get the reward r_t, while other users do not transmit, e.g., TDMA, CSMA with perfect sensing.

Algorithm 3: DLP-A-SCB: Distributed Learning Algorithm with Prioritized Access by A-SCB(D)

1: **Input**: Receive $s_n^m \sim \Upsilon$ for user m. Map the mobile user m locations $\{s_n^m\}_{n\geq 1}$ to a cluster $\mathcal{S}_c$
2: **Initialization**: $[\hat{r}_{i,k,c}]_{N\times K} = [\mathbf{0}]_{N\times K}$ and $[T_{i,c}(n)]_{1\times N} = [\mathbf{0}]_{1\times N}$
3: The BS observes the context $\mathbf{y}_n^m = \mathbf{y}_k^m$ for user m.
4: **for** $n = 1$ to N **do**
5: Allocate channel i such that $i_n = i = ((m+n) \mod N) + 1$;
6: Update $\hat{r}_{i,k,c}^m = r(x_{i_n}, \mathbf{y}_k; c), \forall k : 1 \leq k \leq K$;
7: $T_{i,c}^m(n) = 1$
8: **end for**
9: **while** 1 **do**
10: $n = n + 1$;
11: Allocate a channel D according to the policy A-SCB(D) specified in Alg. 2;
12: Update $\hat{r}_{i,k,c}^m$ and $T_{i,c}^m(n), \forall 1 \leq k \leq K, \forall s_n \in \mathcal{S}_c$ for each user m
$\hat{r}_{i,k,c}^m(n) = \frac{\hat{r}_{i,k,c}^m(n-1)T_{i,c}(n-1) + r(x_{i_n}, \mathbf{y}_k; s_n)}{T_{i,c}^m(n-1)+1}$,
$T_{i,c}^m(n) = T_{i,c}^m(n-1) + 1$
13: **end while**

2) **Collision model** ($\mathbf{M}_2$): No user gets any reward under perfect collision model, e.g., ALOHA protocol.
3) **Multiple reception model** ($\mathbf{M}_3$): All the users get the same reward, e.g., successive interference cancellation (SIC).

If the BS uses round-robin schemes to allow multiple users sharing the channel set witout conflicting, we have

1) **Sharing model** ($\mathbf{M}_1$): A scheduled users j get the reward r_t, while other users will transmit in the rest times, e.g., TDMA.

Note that the contention model and the sharing model have the same resulted reward structure for all users. So, we regard it as the same model in the respective of online learning reward. We denote the three model as $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, respectively.

Denote the policy for each decentralized user j as π_j and the set of policies for all users as $\pi = \{\pi_j, 1 \leq j \leq M\}$. Assuming no information exchange among users, the regret is the gap between the expected reward of all users that could be obtained by a genie-aided optimal allocation and that obtained by the policy π, which can be expressed as

$$\mathcal{R}_\pi(n) = n \sum_{i\in\mathcal{O}_M^*} \mu_{i,\mathbf{y}^i,c} - \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{M} {}_{r_t\sim\nu_{i_t,\mathbf{y}_t,s_t}}[r_t]\,\mathbb{I}_{i,j}(t)\right],$$

where $\mathcal{O}_M^*$ is the set of M channels with M largest expected rewards, and $\sum_{t=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{M} {}_{r_t\sim\nu_{i_t,\mathbf{y}_t,s_t}}[r_t]\,\mathbb{I}_{i,j}(t)$ is the sum of the actual rewards obtained by all users up to slot n. For $\mathbf{M}_1$, $\mathbb{I}_{i,j}(t)$ is defined to be 1 if user j is the one with the smallest index among the users over channel i and 0 otherwise. For $\mathbf{M}_2$, $\mathbb{I}_{i,j}(t)$ is defined to be 1 if user j is the only user over channel i and 0 otherwise. For $\mathbf{M}_3$, $\mathbb{I}_{i,j}(t)$ is defined to be 1 for all users over channel i at time t and 0 otherwise.

As noticed, in the downlink 5G communication scenario, each user will either keep the profile of itself and send its historical information to BS for downlink scheduling ($\mathbf{M}_1$), or keep the profile in its locality ($\mathbf{M}_2$ and $\mathbf{M}_3$). If a user j leave the system, the BS or other users can still utilize its previous records and stop counting the number of rounds and computing its value of UCB any further. Because the UCB-type of algorithm can be implemented in a fully distributed way for all combinatorial problems (e.g. the multi-user case) without the complexity issue [13], no channel will be allocated to the user j, and this will not waste any channel resources or cost any further problems to the remaining users. This property also holds for new arriving users.

In the next, we focus on a prioritized access problem, where it is desired to prioritize a set of ranked users so that the D-th ranked user learns to access the channel with the D-th largest reward. In this scenario, we propose our distributed learning algorithm with prioritized access, DLP-A-SCB, in Alg. 3.

Similar as in Alg. 2, here we need to keep the $N \times K$ vectors $[\hat{r}_{i,k,c}]_{N\times K}$ and the $1 \times N$ vectors $[T_{i,c}(n)]_{1\times N}$ for each user m. We denote o_m^* as the index of channel with the m-th largest expected reward. Note that $\{o_m^*\}_{1\leq m\leq M} = O_M^*$. Since the BS always keeps their accumulated values, therefore the storage complexity is now $\Theta(MNK)$ that does not increase with time, while the computational complexity is upper bounded by $O(MNKt)$. Line 5 ensures that there will be no collision among users. Line 11 in Alg. 3 means the user m is allocated the channel with the D-th largest expected reward with Alg. 2.

Theorem 3. The expected regret $\mathcal{R}_\pi(n)$ under the DLP-A-SCB in Alg. 3 for model $\mathbf{M}_1$ and $\mathbf{M}_2$ is at most

$$\sum_{m=1}^{M}\sum_{s\in S}\left(\sum_{i:i\neq o_{m,s}^*}\bar{T}_{o_m^*,i_n,s_n}(n) + \sum_{h\neq m}\sum_{i\in O_{M,s}^*}\bar{T}_{o_h^*,o_m^*,s_n}(n)\right)\mu_{\max,s}$$

under the definition of $\bar{T}_{o_m^*,i_n,s_n}(n)$ and $\bar{T}_{o_h^*,o_m^*,s_n}(n)$ on $\mathcal{S}$ and is at most

$$\sum_{m=1}^{M}\sum_{c\in\mathcal{C}}\left(\sum_{i:i\neq o_{m,s}^*}\bar{T}_{o_m^*,i_n,\mathcal{S}_{c_n}}(n) + \sum_{h\neq m}\sum_{i\in O_{M,s}^*}\bar{T}_{o_h^*,o_m^*,\mathcal{S}_{c_n}}(n)\right) \times\mu_{\max,c}$$

under the definition of $\bar{T}_{o_m^*,i_n,\mathcal{S}_{c_n}}(n)$ and $\bar{T}_{o_h^*,o_m^*,\mathcal{S}_{c_n}}(n)$ on $\mathcal{C}$; for model $\mathbf{M}_3$ is at most

$$\sum_{m=1}^{M}\sum_{s\in S}\left(\sum_{i:i\neq o_{m,s}^*}\bar{T}_{o_m^*,i_n,s_n}(n) + \sum_{h\neq m}\sum_{i\in O_{M,s}^*}\bar{T}_{o_h^*,o_m^*,s_n}(n)\right)\Delta_{\max,s}$$

on $\mathcal{S}$ and is at most

$$\sum_{m=1}^{M}\sum_{c\in\mathcal{C}}\left(\sum_{i:i\neq o_{m,s}^*}\bar{T}_{o_m^*,i_n,\mathcal{S}_{c_n}}(n) + \sum_{h\neq m}\sum_{i\in O_{M,s}^*}\bar{T}_{o_h^*,o_m^*,\mathcal{S}_{c_n}}(n)\right) \times\Delta_{\max,c}$$

on $\mathcal{C}$, where we have $\bar{T}_{o_m^*,i_n,s_n}(n)$ and $\bar{T}_{o_h^*,o_m^*,s_n}(n)$ on $\mathcal{S}$ by substituting $\Delta_{D_c,i_n,c}^{k_n}$ in the upper bound of regret in (15) with the respective associate values $\Delta_{o_m^*,i_n,c}^{k_n}$ and $\Delta_{o_h^*,o_m^*,c}^{k_n}$, $\bar{T}_{o_m^*,i_n,\mathcal{S}_{c_n}}(n)$ and $\bar{T}_{o_h^*,o_m^*,\mathcal{S}_{c_n}}(n)$ on $\mathcal{C}$ by substituting $\widetilde{\Delta}_{D_C i,c}^{k}$ and $\Delta_{D_c,i_n,c}^{k_n}$ in the upper bound of regret in (16) with respective associate values $\widetilde{\Delta}_{o_m^*,i,c}^{k}$ and $\Delta_{o_h^*,o_m^*,c}^{k_n}$, and $\mu_{max,s} = \max_{1\leq k\leq K, 1\leq i\leq N}\mu_{i,k,s}$ and $\mu_{max,c} = \max_{1\leq k\leq K, 1\leq i\leq N}\mu_{i,k,c}$, and $\Delta_{max,c} = \max_{1\leq k\leq K, 1\leq i\leq N}\max\{\Delta_{o_m^*,i,c}^{k}, \widetilde{\Delta}_{o_m^*,i,c}^{k}\}$ and $\Delta_{max,s} = \max_{1\leq k\leq K, 1\leq i\leq N}\max\{\Delta_{o_m^*,i,s}^{k}, \widetilde{\Delta}_{o_m^*,i,s}^{k}\}$.

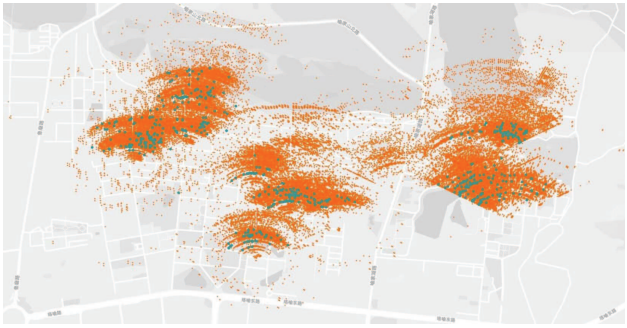


Fig. 2: User locations in the campus of Huazhong University of Science and Technology in China Mobile networks

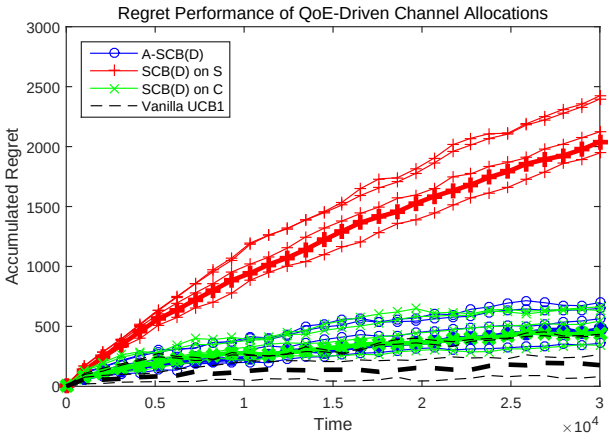


Fig. 3: Regret performance comparison with $N = 20$,$M = 5$,$C = 13$, the best channel in different clusters is *the same*

Proof: See detailed proof in [48]. ■

Remark 2. From the Theorem 3, we find that the regret in the multiple users prioritized setting is upper bounded by $O(MK(N + M - 2)log(n))$. For the fairness consideration in practice, the users should be treated equally important and there should be no priority for the users. Here we can set a new line before the line 11, i.e., $D = ((m+n) \mod M)+1$, such that the D-th largest channel(s) is rotated among the M users, i.e., the users use channels from the estimated largest one to the estimated smallest in turns to ensure the fairness. The analysis of the performance in this fairness setting follows very similar as the proof of Theorem 3 and we omit here for brevity, but we verified its performance in our experiment section.

VII. Experimental Results

In this section, we evaluate the performance of our proposed algorithms for the LTE downlink wireless multimedia communications based on a real collected data, which is provided by China Mobile Communications Corporation. The dataset in our experiments is collected in the region of Wuhan City, located in the Hubei Province, China. This dataset contains of gargantuan description fields such as signal quality and strength of different locations of mobile devices, network information of neighbor plots and etc. Especially, it includes services identifiers for different type of multimedia communications along with fundamental QoS parameters, such as throughput (frame rate (FR)), blocking probabilities, dropping probability (DP), packet delay, PSNR, etc. Although the dateset is offline and contains irregular records, we take some meaningful samples to facilitate the running of our online learning algorithms to verify QoE performance.

A. Dataset Preprocessing and Experiment Setting

This dataset provides a detailed description of network conditions. We extract the useful information about the user location and QoE from the dataset. Based on user locations, we plot the mobility graph of the campus of Huazhong university of science and technology on the Baidu map shown in Fig. 2, which contains a two-week monitoring of all users' records. There are twelve BS covering the whole campus. From the figure, we see that most of the mobility data locations are mainly correct but corrupted by the inaccurate measure of the AOA field. Thus, this is a tough dataset that cause challenging to the clustering algorithms. However, on the other hand, its a good resource to emulate the two typical errors in mobility networks as we indicated at the beginning of Section V.

To perform the single user online learning and multiple user RA online learning for dynamic channel allocation in the LTE network, we collect the communication statistics of each virtual resource block (VRB) with a join operation with service type, e.g., video streaming service from certain different sites, image data transmissions from certain antenna ports, etc.

For QoE metric, based on the study in [42], we devise a novel mean opinion score (MOS) prediction model which incorporates dropping probability (DP), frame rate (FR), peak signal to noise ratio (PSNR), and packet delay variation (PDV) in the dataset. As studied in [43], a typical convention between PDV and MOS is as follows: A constant delay (D) is maintained at $150ms$ and the ΔD as PDV is taken as a measure of the MOS. Only with constant delay $150ms$ and no ΔD i.e., $150ms \pm 0ms$ the value of MOS is high, which obtained for both fast and slow moving videos are more than perceptible and are not annoying. When $D \pm \Delta$ equals to $150 \pm 0ms$, $150 \pm 10ms$, $150 \pm 25ms$, $150 \pm 75ms$ and $150 \pm 100ms$, the MOS ranges from $4 \sim 4.15$, $3.41 \sim 4.19$, $3.31 \sim 4.19$, $2.92 \sim 4.08$ and $2.5 \sim 3.31$ for several type of TCP and UDP protocols. Another example is described in [42] for the conversion between PSNR and MOS. In the LTE network, we have the values of PSNR $(dB) > 37.67$, $32.00 \sim 37.01$, $26.00 \sim 31.21$, $22.00 \sim 25.91$, < 21.86 corresponds to MOS values "5, 4, 3, 2, 1", respectively. Based on the study in [42], we devise a novel MOS prediction model which incorporates the DP, FR, PSNR, and PDV into the MOS calculation as shown below:

$$MOS = \frac{e_1 + e_2 FR + e_3 \ln(PSNR)}{1 + e_4 DP + e_5 (DP)^2 + e_6 PDV}, \quad (17)$$

where the regression coefficient e_1 to e_6 is unknown in the prediction model. In our experiments, we use the matlab nonlinear toolbox to perform a nonlinear regression analysis of the dataset offline, where the coefficients are $e_1 = 8.3025, e_2 = -0.9371, e_3 = 0.5723, e_4 = 1.9612, e_5 = 9.10246 and e_6 =$

0.1238 along with R^2 showing the goodness of fit for the video applications over the LTE networks. The model is verified with three different video sequences of suzie, carphone and football in the three corresponding slow, median and fasting moving content categories. MATLAB$^{\text{TM}}$ function `nlintool` has been used to carry out the nonlinear regression analysis. R^2 indicates the goodness of fit of the fitted coefficients of the models, which is on average of 92.31% in our model. In our experiments, the set $\{FR, PSNR, DP, PDV\}$ is the context set for the online learning algorithms.

B. Regret Performance Comparison

We conduct our experiments under different groups with numbers of channels N, users M and clusters C, and compare the performance of algorithms A-SCB(D), SCB(D) on S, SCB(D) on C and Vanilla UCB1. A-SCB(D) is our proposed practical agnostic algorithm. In contrast, SCB(D) on S is the ideal and raw version of our algorithm that does not explore the diversity of user location information, and SCB(D) on C is the ideal and theoretical version of our algorithm that assumes known priori information of all clusters and the cluster type of all users' locations. Moreover, Vanilla UCB1 represents all the previous MAB algorithms (e.g., [31]–[37]) that do not utilize the QoE context and user behavioral information. Fig. 3 and Fig. 4 show the regrets comparisons of the proposed algorithms in different scenarios. The vanilla UCB1 [11] and SCB(D) on $\mathcal{S}$ are adopted as baselines. Note that we adopt the prioritized multi-user RA online learning for all the algorithms, e.g. the A-SCB(D) is actually DLP-A-SCB(D). The thick lines in the figures are used to represent the mean regrets and the dashed lines for quantiles at levels $0.25, 0.5, 0.75, 0.95$ and 0.99. By sampling the dataset, we set the distortion factors γ_c as 3.24. The algorithms run over 100 trials for a large horizon $n = 30000$, where the QoE performance metrics shown in the next subsection are satisfactory.

Fig. 3 presents a scenario where the best channel in different clusters is the same and the distribution of CSI among different channels are quite close. In this case, it is very hard for the A-SCB(D) to distinguish different channels (to find the D-optimal ones) and different clusters. We find that the SCB(D) on $\mathcal{S}$ performs poorly, and the A-SCB(D) is defeated even by UCB1 slightly, because there is one channel as the best for all contexts in this scenario. This indicates our proposed algorithms should have potential to work well in the more complex dynamic changing wireless environments.

Fig. 4(a) presents an expected performance, where both the vanilla UCB1 and the SCB(D) on $\mathcal{S}$ perform poorly with respect to the offline optimal value. In this scenario, the best channel is different in the different clusters, and the corresponding value is always very high and are well separated from other channels by A-SCB(D). Comparing to the A-SCB(D), the vanilla UCB1 without human behavior and QoE data included, has an 228.73% value of regret increase. This indicates great performance gain of A-SCB(D) in this scenario, and A-SCB(D) is especially good at differentiating *channels with different qualities*, and predict the best CSI for DCA in wireless communications over time.

TABLE II: Blocking Probability

(M, C)	Algorithm	**Probability** \ **round** 3×10^4 $N = 8$	$N = 16$	$N = 24$	$N = 32$	$N = 48$	$N = 64$
(50, 74)	**A-SCB(D)**	.561	.534	.512	.483	.478	.452
	SCB(D) on S	.675	.643	.592	.577	.569	.547
	SCB(D) on C	.509	.481	.461	.435	.430	.407
	Vanilla UCB1	.622	.602	.574	.538	.532	.525
	No Learn	.729	.684	.666	.618	.621	.588
(20, 74)	**A-SCB(D)**	.621	.584	.543	.513	.491	.467
	SCB(D) on S	.732	.689	.641	.605	.578	.551
	SCB(D) on C	.571	.537	.532	.461	.447	.425
	Vanilla UCB1	.745	.701	.652	.616	.589	.560
	No Learn	.776	.730	.678	.646	.624	.592

TABLE III: Throughput (FR)

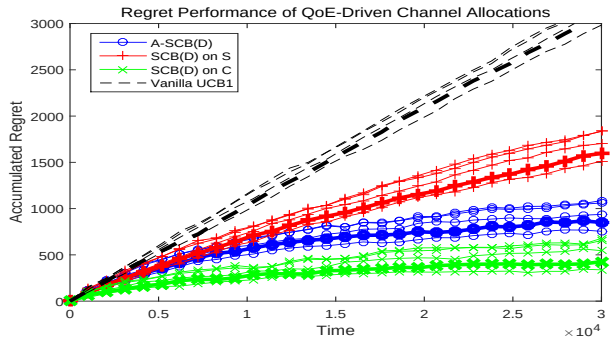
(M, C)	Algorithm	**Kbps** \ **round** 3×10^4 $N = 8$	$N = 16$	$N = 24$	$N = 32$	$N = 48$	$N = 64$
(50, 74)	**A-SCB(D)**	61.3	82.4	104.6	106.1	109.3	113.3
	SCB(D) on S	52.2	70.1	88.4	90.1	92.9	96.3
	SCB(D) on C	73.2	98.8	124.8	127.2	131.2	136.1
	Vanilla UCB1	54.9	74.2	94.1	95.5	98.1	.102.0
	No Learn	49.0	65.9	83.7	84.9	87.2	90.6
(20, 74)	**A-SCB(D)**	55.1	69.1	74.2	85.6	90.4	93.2
	SCB(D) on S	46.8	58.7	62.9	72.8	76.8	79.2
	SCB(D) on C	66.0	82.8	89.0	102.7	108.1	111.8
	Vanilla UCB1	49.5	62.1	66.8	77.0	81.4	83.9
	No Learn	44.0	55.2	59.4	68.5	72.3	74.4

Fig. 4(b) presents a variant on a dataset when N is large. As expected, the performance of all algorithms degrades, but A-SCB(D) is still competitive with respect to the baseline SCB(D) on $\mathcal{S}$. In this scenario, comparing to the A-SCB(D), the vanilla UCB1 has an 113.09% value of regret increase. With larger number of channels, the A-SCB(D) is more capable to explore the *multi-channel diversity* than vanilla UCB1 in wireless communications, which proves its advantage and effectiveness.

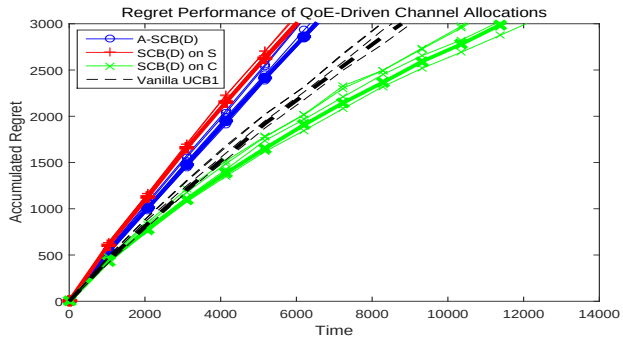
Fig. 4(c) presents a variant on a dataset when M is large, one user only dwells on each location s 10 times less than the previous dataset, which is challenging. From the results, we find that the A-SCB(D) behaviors initially like SCB(D) on S, and gradually it behaviors like SCB(D) on $\mathcal{C}$. In this scenario, comparing to the A-SCB(D), the vanilla UCB1 has an 517.18% value of regret increase. This inundates that the proposed A-SCB(D) performs extremely well in exploring the *multi-user and mobility diversity*, which has great potential to be applied in the large-scale and high mobility network conditions.

Fig. 4(d) presents a variant on a dataset when $\mathcal{C}$ is large and the number of users is also large. Although we find that most of the users are within $10 \sim 30$ number of clusters, we see that A-SCB(D) still works fairly decent in this case.

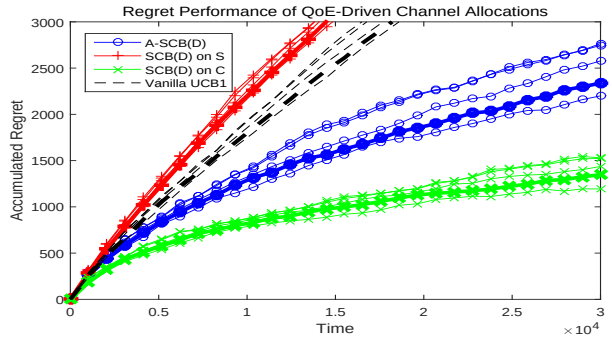
As a summary of results in Fig. 4, A-SCB(D) consistently competes with SCB(D) on $\mathcal{C}$, while UCB1 and SCB(D) on $\mathcal{S}$ obtain poor regrets in typical and practical scenarios. This indicates that our proposed algorithm A-SCB(D) by incorporating human-behavioral data into channel allocation can greatly improve the performance of mobile communications.



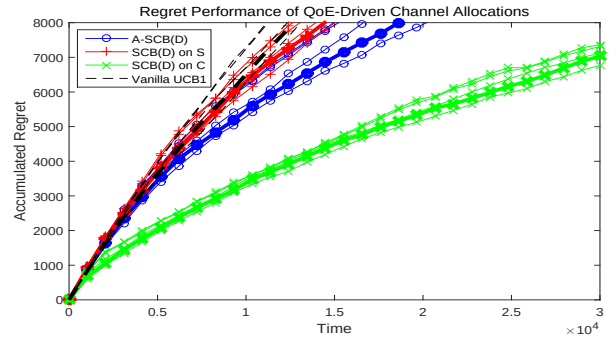
(a) $N = 20$, $M = 5$, $C = 13$



(b) $N = 64$, $M = 5$, $C = 13$



(c) $N = 32$, $M = 50$, $C = 74$



(d) $N = 32$, $M = 20$, $C = 74$

Fig. 4: Regret performance comparison where the best channel in different clusters is *different*

TABLE IV: Dropping Probability

(M, C)	**Algorithm**	**Probability \ round** 3×10^4 $N = 8$	$N = 16$	$N = 24$	$N = 32$	$N = 48$	$N = 64$
(50, 74)	**A-SCB**(D)	.061	.067	.072	.086	.109	.123
	SCB(D) **on** S	.0738	.081	.086	.104	.131	.148
	SCB(D) **on** C	.055	.060	.066	.077	.099	.110
	Vanilla UCB1	.072	.079	.085	.101	.129	.145
	No Learn	.076	.084	.090	.107	.136	.154
(20, 74)	**A-SCB**(D)	.061	.068	.075	.089	.111	.124
	SCB(D) **on** S	.072	.080	.089	.105	.130	.146
	SCB(D) **on** C	.058	.065	.071	.084	.105	.118
	Vanilla UCB1	.074	.082	.091	.108	.134	.150
	No Learn	.076	.085	.092	111	.167	.187

TABLE V: Measured Average PDV

(M, C)	**Algorithm**	**ms \ round** 3×10^4 $N = 8$	$N = 16$	$N = 24$	$N = 32$	$N = 48$	$N = 64$
(50, 74)	**A-SCB**(D)	26.1	23.4	18.6	15.4	14.3	12.2
	SCB(D) **on** S	31.2	28.1	22.3	18.5	17.2	14.6
	SCB(D) **on** C	23.4	21.1	16.7	15.6	15.1	14.2
	Vanilla UCB1	30.7	27.1	21.0	17.6	16.1	14.9
	No Learn	32.7	28.9	23.1	19.7	17.7	15.6
(20, 74)	**A-SCB**(D)	27.7	25.3	21.3	18.2	15.7	13.9
	SCB(D) **on** S	30.47	27.9	23.1	20.1	16.8	14.2
	SCB(D) **on** C	24.9	22.3	18.4	16.3 15.4	13.4	
	Vanilla UCB1	31.3	28.1	24.4	21.3	17.8	15.2
	No Learn	36.1	32.4	27.1	23.0	18.6	16.6

C. MOS Performance Metrics

We also compare the performance of obtained results of blocking probability, FR, DP, PDV and PSNR, under the QoE performance metric of the MOS for the algorithms A-SCB(D), SCB(D) on S, SCB(D) on C, Vanilla UCB1 and the plain case where no machine learning algorithm is applied ("No Learn"). We have listed the respective comparison results in Table II, Table III, Table IV, Table V and Table VI under two typical (M, C) user-context pairs $(50, 74)$ and $(20, 74)$ under different size of channels N.

In Table II, we can find that the A-SCB(D) have a 9%-13% blocking probability reduction than Vanilla UCB1 and a reduction of 13%-29% of blocking probability than "No Learn" on average for all the scenarios. In Table III, we can find that the A-SCB(D) have a 18%-26% throughput improvement than Vanilla UCB1 and a improvement of 22%-34% of FR than "No Learn" on average for all the scenarios. In Table IV, we can find that the A-SCB(D) have a 11%-17% DP reduction than Vanilla UCB1 and a reduction of 19%-31% of DP than "No Learn" on average for all the scenarios. In Table V, we can find that the A-SCB(D) have a 27%-35% PDV reduction than Vanilla UCB1 and a reduction of 36%-54% of PDV than "No Learn" on average for all the scenarios .In Table VI, we can find that the A-SCB(D) have a around $1.8dB$ PSNR improvement than Vanilla UCB1 and a reduction of $3.2dB$ PSNR improvement than "No Learn" on average for all the scenarios.

In summary, when we transfer the values of FR, DP, PDV and PSNR into the reward of MOS as the QoE performance. We figure out that on average the our proposed A-SCB(D) have a range of $12\% - 44\%$ QoE improvement than Vanilla UCB1 and a range of $25\% - 57\%$ QoE improvement than "No Learn" cases. This strongly support that our proposed algorithms have great capability to improve the QoE performance

TABLE VI: PSNR

(M, C)	Algorithm	***dB*** $\setminus$ **round** 7×10^7				
		$N = 12$	$N = 24$	$N = 32$	$N = 48$	$N = 64$
(50, 74)	**A-SCB**(D)	36.29133	36.53665	37.20356	37.96044	39.19218
	SCB(D) **on** S	33.76521	34.01298	36.39812	36.89234	38.24134
	SCB(D) **on** C	38.56413	39.12341	42.13119	43.09881	45.23142
	Vanilla UCB1	34.33315	35.75614	36.89124	37.34213	38.54146
	No Learn	33.01298	33.68431	34.87545	35.41490	36.56721
(20, 74)	**A-SCB**(D)	32.12351	32.87234	33.34986	34.54129	36.12963
	SCB(D) **on** S	30.91234	30.35421	31.12412	32.98872	35.67512
	SCB(D) **on** C	34.51861	34.31985	35.99051	36.87512	39.02195
	Vanilla UCB1	30.20297	30.98513	31.98155	33.20124	35.97851
	No Learn	29.28213	29.81431	30.41512	32.51342	33.90818

in future wireless communications.

D. Performance of Prioritized and Fair Learning Algorithms

Next, we conduct the experiments in the the multi-user RA scenario. Without loss of generality, we compare the specific performance indices terms under the MOS for both prioritized and fair channel access in the MAC scheme $\mathbf{M}_3$ under SIC. Fig. 5 provides the offline optimal values for the prioritized access ("prio opt"), prioritized access by the proposed online learning algorithm ("prio learn") and fair access by the proposed online learning algorithm ("fair learn"). In the "prio learn" scheme, we consider three EUs, in which EU 1 has the highest priority, EU 2 has the medium priority and EU 3 has the lowest priority.

Fig. 5(a) presents the blocking probability of EUs, which describes the number of blocked new connections. The results of the blocking probability verify the theoretical analysis for both schemes. Moreover, the "prio learn" achieves a very quick convergence to its offline optimal value, which demonstrates the accuracy and correctness of our proposed algorithm. In addition, the performance difference between the "prio learn" and "fair learn" schemes is not significant. In Fig. 5(b), we present the dropping probability of EUs, which describe the number packet drops in wireless multimedia communications. EU 1 has the least values of DP, while EU 3 has the highest values of DP in the "prio learn" scheme. However, the DPs of all EUs are close to each other in the "fair learn" scheme. Fig. 5(c) presents the throughput performance of EUs. The proposed online learning algorithm can always guarantee that the EU with a higher priority have a higher throughput, while there is a tradeoff on DP performance among multiple users with the increase of number of channels in Fig. 5(b). Fig. 5(d) presents the delay performance of the EUs. The values of PDV for both 'prio learn" and 'fair learn" schemes are ranged from $5 \sim 20$ms, which correspond to acceptable MOS for the QoE performance. For example, EU 1 with a PDV value around 5 indicates a very high QoE for the user with the highest priority. In reality, some clusters may be more important, e.g., running more applications with more stringent QoE requirements. For the number of prioritized users who have better QoE performance, we could identify them to be the more prioritized clusters, which are expected to have better channel conditions and offer better QoE-aware services. This information could be collected for the wireless communication environment planning, and we can put the more important users within these prioritized clusters.

Furthermore, we study the real video transmission by three groups of datasets, which includes the fast moving videos, median moving videos and slow moving videos. We focus on the PSNR performance that is statistically calculated by the video streaming protocols and mapped it on each corresponding channel. We list the measured average video PSNR in Table VII. From the multiple user distributed learning, EU 1 always achieves a higher PSNR than other two EUs, and EU 1 has the lowest PSNR. With the increase of number of channels, the searching space increase for seeking the video frames with larger PSNR values, and thus the PSNR values of all EUs increase. By the mapping from PSNR to MOS, the proposed channel allocation scheme can achieve the good multimedia transmission quality for all users. Specifically, the average QoE improvement for 'prio learn" over the 'fair learn" scheme is about 8%-17%. This indicates that providing priority services to prioritized users could effective improve their QoE performance in our algorithm setting.

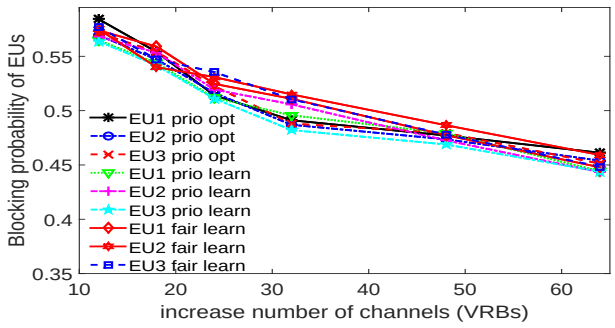
VIII. Conclusion and Future Work

We design effective online contextual bandit learning algorithms to facilitate QoE-driven dynamic channel allocation by incorporating human behavioral data for 5G networks. We propose an agnostic Latent SCB algorithm to resolve this hard problem. Specifically, our proposed A-SCB(D) algorithm has an about 30% QoE improvement than classic Vanilla UCB1 algorithm and an about 45% QoE improvement than the case without implementing machine learning algorithm. Moreover, the prioritized version of our algorithm shows an additional 10% QoE improvement over the fairness learning version. This demonstrates that exploring human behavioral data indeed has great potential to improve the performance of 5G networks.

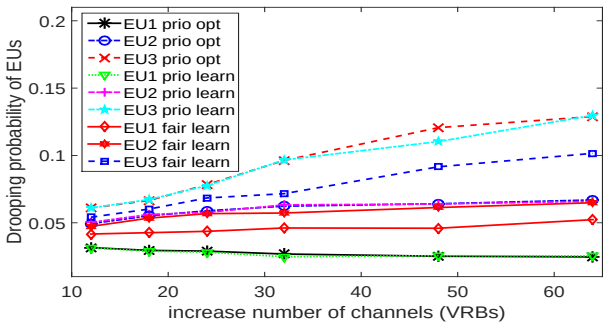
For future research, to further embrace the power of big data analytics for DCA in 5G wireless communications, we strongly suggest to devise the low-complexity deep reinforcement learning algorithms to model trajectory prediction and radio maps by considering the spatial channel distribution variation. As we can see, the curse of dimensionality issue hinders its progress, which must be heavily resolved at first. This is especially true for the model-free reinforcement learning algorithms. To resolve this problem, we suggest to study the model-based reinforcement learning in a latent space [29], which is very promising solution. The idea is quite similar to the LCAMB proposed in this paper, which uses the latent representation to discard information freely that is irrelevant to the design goal. Plus, we can utilize the idea of the hierarchical or tree-based structure to further reduce the complexity of learning algorithm, such as the one in [30].

Acknowledgement

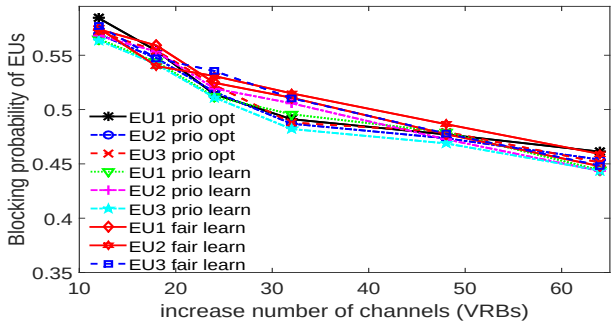
The work of Pan Zhou was supported by the National Natural Science Foundation of China under Grant No. 61972448. The work of Changkun Jiang was supported by the National Natural Science Foundation of China under Grant No. 61902255, the Startup Foundation of Shenzhen under Grant



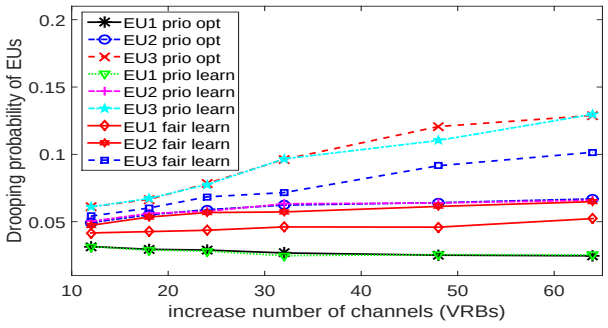
(a) Blocking probability



(b) Dropping probability



(c) Throughput



(d) Measured averaged PDV

Fig. 5: Performance of EUs under MOS as QoE metric

TABLE VII: The Measured Average Video PSNR with Increase Number of Channels

Type / Rate	EU 1			EU 2			EU 3		
Increase Number of Channels	fast video	median video	slow video	fast video	median video	slow video	fast video	median video	slow video
12	36.29133	36.65173	37.65939	35.36165	35.64346	36.19088	34.47124	35.63634	35.87170
24	36.53665	36.99012	38.09187	36.01525	36.21356	36.85125	35.31135	36.91717	36.13174
32	37.20356	37.59135	38.89033	36.21376	37.07171	37.41124	35.97614	37.24383	36.53157
48	37.96044	38.24769	39.27915	36.96077	37.86118	38.77616	36.26713	37.84185	37.73105
64	38.19218	39.08355	39.84345	37.72937	38.22178	39.03262	36.72347	37.88082	38.43155

No. 827-000415, and the Natural Science Foundation of SZU under Grant No. 860-000002110540. The work of Pan Zhou was supported by Zhejiang Provincial Natural Science Foundation of China (No. LR17F010001). The work of Wei Wang was supported by the National Natural Science Foundation of China under Grant No. 61672395.

REFERENCES

[1] C. She and C. Yang, "Context aware energy efficient optimization for video on-demand service over wireless networks," in *Proc. IEEE ICCC*, pp. 1–6, Nov. 2015.

[2] A. Nadembega, A. Hafid, and T. Taleb, "Mobility-prediction-aware bandwidth reservation scheme for mobile networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2561–2576, Jun. 2015.

[3] H. Abou-Zeid and H. S. Hassanein, "Toward green media delivery: Location-aware opportunities and approaches," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 38–46, Aug. 2014.

[4] L. Duan, L. Huang, C. Langbort, A. Pozdnukhov, et al, "Human-in-the-Loop Mobile Networks: A Survey of Recent Advancements," *IEEE J. S. Areas in Comm.*, vol. 35, no. 4, pp. 813-831, Apr. 2017.

[5] D. S. Nunes, P. Zhang, and J. S. Silva, "A survey on human-in-the-loop applications towards an internet of all," *IEEE Commun. Surv. Tut.*, vol. 17, no. 2, pp. 944-965, 2015.

[6] K. U. R. Laghari and K. Connelly, "Toward total quality of experience: A QoE model in a communication ecosystem," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 58-65, Apr. 2012.

[7] X. Chen, F. Mériaux, S. Valentin,"Predicting a user's next cell with supervised learning based on channel states," *Proc. of IEEE SPAWC*, pp. 36-40, Jun 16, 2013.

[8] J. Langford and T. Zhang, " The epoch-greedy algorithm for multi-armed bandits with side information," *Proc. of NIPS*, Dec. 2007.

[9] T. Lu, D. Pal, and M. Pal, "Contextual multi-armed bandits," *Proc. of AISTATS*, May 2010.

[10] P. Sakulkar, B. Krishnamachari, "Stochastic contextual bandits with known reward functions," arXiv:1605.00176, 2016.

[11] P. Auer, N. Cesa-Bianchi and P. Fischer, "Finite-time analysis of the multi-armed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235-256, May 2002.

[12] O. Maillard and S. Mannor, "Latent bandits," *Proc. of IEEE ICML*, Jun. 2014.

[13] J. Y. Audibert, S. Bubeck, G. Lugosi, "Regret in online combinatorial optimization," Mathematics of Operations Research, vol. 39, no. 1, pp. 31-45, May 6, 2013.

[14] R. Margolies et al., "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 355–367, Feb. 2016.

[15] N. Bui and J. Widmer, "Mobile network resource optimization under imperfect prediction," in *Proc. IEEE WoWMoM*, Jun. 2015, pp. 1–9.

[16] R. Atawia, H. Abou-zeid, H. S. Hassanein, and A. Noureldin, "Robust

resource allocation for predictive video streaming under channel uncertainty," in *Proc. IEEE GLOBECOM*, Dec. 2014, pp. 4683–4688.
[17] C. Luo, J. Ji, Q. Wang, X. Chen, P. Li, "Channel state information prediction for 5G wireless communications: A deep learning approach," *IEEE Trans. on Net. Science and Eng.*, Jun 25, 2018.
[18] Y. Sui, W. Yu, and Q. Luo,"Jointly Optimized Extreme Learning Machine for Short-Term Prediction of Fading Channel, "*IEEE Access*, vol.6, pp.49029-49039, 2018.
[19] Q. Wu, Z. Du, P. Yang, Y. D. Yao, J. Wang, "Traffic-Aware Online Network Selection in Heterogeneous Wireless Networks,"*IEEE Trans. Vehicular Tech.*, vol.65, no.1, pp. 381-97, Jan., 2016.
[20] R. Annavajjala, R. S. Mangoubi, C. Y. Christopher, J. M. Zagami, "An online learning approach to throughput optimization in wireless networks under dynamic and unknown interference conditions," *In Proc. of IEEE MLSP*, pp. 1-6, Sep 17, 2015.
[21] P. Charonyktakis, M. Plakia, I. Tsamardinos, M. Papadopouli, "On user-centric modular qoe prediction for voip based on machine-learning algorithms," *IEEE Trans. on mobile comput.* , vol.15, no.6, pp.1443-56, Jun 1, 2016.
[22] Y. T. Lin, E. M. Oliveira, S. B. Jemaa ,S. E. Elayoubi, "Machine learning for predicting QoE of video streaming in mobile networks," *In Proc. of IEEE ICC*, pp. 1-6, May 21, 2017.
[23] X. Chen,F. Mériaux, S. Valentin,"Predicting a user's next cell with supervised learning based on channel states," *In Proc. of IEEE SPAWC*, pp. 36-40, Jun 16, 2013.
[24] C. Yao, C. Yang, Z. Xiong,"Energy-Saving Predictive Resource Planning and Allocation," *IEEE Trans. Comm.*, vol.64, no.12, pp.5078-95, 2016
[25] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. C. Liang, D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133-3174, 2019.
[26] R. S. Sutton and A. G. Barto, Introduction to reinforcement learning, vol. 2, no. 4., *Cambridge: MIT press*, 1998.
[27] Y. GullapaUi, "A stochastic reinforcement learning algorithm for learning real-valued functions," *Neural Networks*, vol. 3, no. 6, pp. 671 - 692, 1990.
[28] S. Koenig,R. G. Simmons, "Complexity analysis of real-time reinforcement learning,"*In proc. of AAAI*, pp. 99-107, Jul. 11, 1993.
[29] A. X. Lee, A. Nagabandi, P. Abbeel, S. Levine, "Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model," arXiv preprint arXiv:1907.00953, Jul 1, 2019.
[30] T. Haarnoja, K. Hartikainen, P. Abbeel, S. Levine,"Latent space policies for hierarchical reinforcement learning," *In proc. of ICML*, pp. 1-10, Apr 9., 2018.
[31] K. Wang and L. Chen, "On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach," *IEEE Trans. Sig. Process.*, vol. 60, no. 1, pp. 300-309, Jan. 2012.
[32] K. Liu, Q. Zhao, B. Krishnamachari, "Dynamic multichannel access with imperfect channel state detection," *IEEE Trans. Sig. Process.*, vol. 58, no. 5, pp. 2795-2808, Jan. 2012.
[33] Y. Gai and B. Krishnamachari, "Distributed stochastic online learning policies for opportunistic spectrum access," *IEEE Trans. Sig. Process.*, vol. 62, no. 23, pp. 6184-6193, Dec. 2014.
[34] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Sig. Process.*, vol. 58, no. 11, pp. 5667-5681, Nov. 2010.
[35] N. Modi, P. Mary, C. Moy, 'QoS Driven Channel Selection Algorithm for Cognitive Radio Network: Multi-User Multi-Armed Bandit Approach," *IEEE Trans. Cogn. Comm. and Net.*, vol.3, no. 1, pp.49-66, Mar 1, 2017.
[36] A. Anandkumar, N. Michael, and A. Tang, "Distributed learning and allocation of cognitive users with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 781-745, Apr. 2011.
[37] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," *Proc. of IEEE DySPAN*, Apr. 2010.
[38] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," *Proc. of ACM WWW*, Apr., 2010.
[39] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a predictive model of quality of experience for internet video," *Proc. of ACM SIGCOMM*, Aug. 2013.
[40] W. Wanalertlak, B. Lee, C. Yu, S.-M. Park, and W.-T. Kim, "Behavior-based mobility prediction for seamless handoffs in mobile wireless networks," *Wireless Networks*, vol. 17, no. 3, pp. 645-658, Apr. 2011.
[41] Y. Bao, H. Wu, and X. Liu, "From prediction to action: A closed-loop approach for data-guided network resource allocation," *Proc. ACM KDD*, Aug. 2016.
[42] A. Khan, L. Sun, E. Jammeh and E. Ifeachor, "Quality of experience driven adaptation scheme for video applications over wireless networks," *IET Commun*, vol. 4, no. 11, pp. 1337-1347, Nov. 2011.
[43] P. Uppu, and S. Kadimpati, "QoE of Video Streaming over LTE Network", https://www.diva-portal.org/smash/get/diva2:829766/FULLTEXT01.pdf, master thesis, 2013.
[44] A. Slivkins, "Contextual bandits with similarity information," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2533-2568, 2014.
[45] S. Boucheron, G. Lugosi, P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.
[46] V. V. , Vazirani, Approximation algorithms, Springer Science & Business Media, 2013.
[47] LTE Technical Specification. # 36.211 - 3GPP, www.3gpp.org/dynareport/36211.htm
[48] Pan Zhou, et al, "Human-Behavior and QoE-Aware Dynamic Channel Allocation for 5G Networks: A Latent Contextual Bandit Learning Approach," *technical report*, https://www.dropbox.com/s/3lz35fkhsbcysl8/TCCN19TR.pdf?dl=0, July, 2019.

Pan Zhou(S'07–M'14) is currently an associate professor with School of Cyber Science and Engineering, Wuhan, P.R. China. He received his Ph.D. in the School of Electrical and Computer Engineering at the Georgia Institute of Technology (Georgia Tech) in 2011, Atlanta, USA. He received his B.S. degree in the Advanced Class of HUST, and a M.S. degree in the Department of Electronics and Information Engineering from HUST, Wuhan, China, in 2006 and 2008, respectively. He held honorary degree in his bachelor and merit research award of HUST in his master study. He was a senior technical member at Oracle Inc., America, during 2011 to 2013, and worked on Hadoop and distributed storage system for big data analytics at Oracle Cloud Platform. He received the "Rising Star in Science and Technology of HUST" in 2017. His current research interest includes: security and privacy, big data analytics and machine learning, and information networks.

Jie Xu (M'15) is an assistant professor in the Department of Electrical and Computer Engineering at the University of Miami, FL, USA. He received B.S. and M.S. in Electronic Engineering from Tsinghua University, in 2008 and 2010, respectively, and Ph.D. in Electrical Engineering from University of California Los Angeles (UCLA) in 2015. He is a recipient of the distinguished Ph.D. dissertation award from UCLA and a recipient of the best paper award at APCC. His research interests include mobile edge computing/caching, green communications and network security.

Wei Wang (S'08-M'10-SM'15) received the B.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications, China, in 2004 and 2009, respectively. He is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, China. His research interests mainly focus on stochastic optimization for cross-layer resource allocation, and caching and computing in wireless networks.

Changkun Jiang (S'14-M'10-SM'17) received the PhD degree in information engineering from the Chinese University of Hong Kong in 2017. His research interests include dynamic pricing and revenue management in communication networks, game theory and incentive mechanism design in network economics, and network optimization. He is a member of the IEEE.

Kehao Wang received the BS degree in Electrical Engineering, MS degree in Communication and Information System from Wuhan University of Technology, Wuhan, China, in 2003 and 2006, respectively, and Ph.D in the Department of Computer Science, the University of Paris-Sud XI, Orsay, France, in 2012. From Feb. 2013 to Aug. 2013, he was a postdoc with the HongKong Polytechnic University. In 2013, he joined the School of Information Engineering at the Wuhan University of Technology, where he is currently an associate professor. From Dec. 2015, he has been a visiting scholar in the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA. His research interests include: stochastic optimization, operation research, scheduling, wireless network communications, and embedded operating system.

Jia Hu received the B.E. and M.E. degrees in communicationengineeringandphysicalelectronics from the Huazhong University of Science and Technology, Wuhan, China, in 2004 and 2006, respectively, and the Ph.D. degree in computing from the University of Bradford, U.K., in 2010. He is currently a Lecturer with the Department of Computer Science, University of Exeter, U.K. His research interests include performance modeling and analysis, network protocols and algorithms, next generation networks, cross-layer optimization, network security, and resource management.