Multi-User Small Base Station Association via Contextual Combinatorial Volatile Bandits

Muhammad Anjum Qureshi, *Member, IEEE*, Andi Nika, and Cem Tekin, *Senior Member, IEEE*

***Abstract*— We propose an efficient mobility management solution to the problem of assigning small base stations (SBSs) to multiple mobile data users in a heterogeneous setting. We formalize the problem using a novel sequential decision-making model named contextual combinatorial volatile multi-armed bandits (MABs), in which each association is considered as an arm, volatility of an arm is imposed by the dynamic arrivals of the users, and context is the additional information linked with the user and the SBS such as user/SBS distance and the transmission frequency. As the next-generation communications are envisioned to take place over highly dynamic links such as the millimeter wave (mmWave) frequency band, we consider the association problem over an unknown channel distribution with a limited feedback in the form of acknowledgments and under the absence of channel state information (CSI). As the links are unknown and dynamically varying, the assignment problem cannot be solved offline. Thus, we propose an online algorithm which is able to solve the user-SBS association problem in a multi-user and time-varying environment, where the number of users dynamically varies over time. Our algorithm strikes the balance between exploration and exploitation and achieves sublinear in time regret with an optimal dependence on the problem structure and the dynamics of user arrivals and departures. In addition, we demonstrate via numerical experiments that our algorithm achieves significant performance gains compared to several benchmark algorithms.**

***Index Terms*— Small base stations, dynamic user association, contextual bandits, volatile bandits.**

I. Introduction

WHILE small base stations (SBSs) are envisioned to provide resilient services to pervasive mobile devices in dense cellular networks, their success crucially depends on solving new challenges associated with network load management. As a stepping stone, the case when multiple users pursue connection to an SBS in order to maximize their quality of service (e.g., throughput) has been thoroughly investigated in prior works [1], [2]. On the other hand, when there are multiple SBSs present in the network, the association problem becomes a combinatorial assignment problem. When channel state information (CSI) is given and the users are static, the user-SBS association problem can be solved by Hungarian algorithm or its variants [3], [4]. Unfortunately, this is not possible in practice due to the following reasons. First of all, rapidly varying nature of next-generation communication bands (e.g., millimeter wave (mmWave)) makes it impossible to acquire and utilize CSI efficiently. Secondly, dynamic and unpredictable mobility patterns of the users make static user-SBS associations highly suboptimal. Therefore, efficient deployment of SBSs requires development of new learning methods that will: (i) handle the lack of CSI and accurate channel statistics by learning from past experience; (ii) enable adaptation of user-SBS associations based on contextual information regarding the mobility of the users. In particular, the performance associated with a given user-SBS association depends on many factors such as user/SBS distance and transmission frequency, and consequently, this additional contextual information needs to be taken into account when computing the optimal association in every round.

Manuscript received September 5, 2020; revised November 19, 2020 and January 20, 2021; accepted February 28, 2021. Date of publication March 9, 2021; date of current version June 16, 2021. The work of Cem Tekin was supported by the BAGEP Award of the Science Academy. The associate editor coordinating the review of this article and approving it for publication was C.-H. Lee. *(Muhammad Anjum Qureshi and Andi Nika contributed equally to this work.) (Corresponding author: Muhammad Anjum Qureshi.)*

The authors are with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: qureshi@ee.bilkent.edu.tr; andi.nika@bilkent.edu.tr; cemtekin@ee.bilkent.edu.tr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2021.3064939.

Digital Object Identifier 10.1109/TCOMM.2021.3064939

A. Contributions

In this paper, we propose a solution to the user-SBS association problem with dynamically changing user presence using a new reinforcement learning method called contextual combinatorial volatile multi-armed bandit (MAB). Our setup can handle dynamically arriving/departing users with any mobility pattern. Our algorithm is able to learn the optimal assignment in each round even when the channel conditions are rapidly varying and unknown. The main contributions of this paper are summarized as follows.

- We consider the problem of user-SBS association under dynamic users over rapidly varying channels with unknown statistics and formalize it as a contextual combinatorial volatile MAB problem.
- We propose a combinatorial learning algorithm named MUSIC for the aforementioned problem that exploits the contextual information and volatility of the users. MUSIC is able to achieve performance very close to that of an oracle benchmark that selects the best association in every round. This hypothetical benchmark is impossible to implement in practice since the channel conditions, and consequently, the best association is unknown in each round. We characterize the loss of MUSIC with respect to the oracle benchmark by using the notion of regret, and prove that MUSIC is an almost optimal learning algorithm that achieves $\tilde{O}(T^{(\tilde{D}+1)/(\tilde{D}+2)+\epsilon})$ regret for any $\epsilon > 0$, where $\tilde{D}$ represents the approximate-optimality dimension related to the space of contexts. To the best of our knowledge, this is the first work that proposes a prov-

ably optimal learning algorithm for the dynamic user-SBS association problem in an unknown environment.

- Experimental results demonstrate that MUSIC is able to optimally solve the dynamic association problem under various mmWave channel and user mobility models, outperforming the state-of-the-art by a significant margin.

B. Comparison With Related Works

Most prior works on OFDMA resource allocation solve subcarrier assignment along with the power allocation over these subcarriers [9], [10]. However, it is shown in earlier works [11], [12] that an equal power allocation yields performance very close to optimal. Therefore, the assignment problem appears to be more challenging than power allocation [13]. Since each SBS can only serve a limited number of users, the user association problem in a small cell network is much more challenging than the typical classical cellular user association problem [6]. In this work, we integrate additional constraints to this problem by introducing user mobility and volatility, and the absence of CSI, which is a common constraint under rapidly varying wireless channel conditions. It is important to note that the network's goal is to maximize the overall throughput by choosing the user-SBS associations whose overall sum in terms of the QoS is maximum, a contradicting goal compared to a single user's goal, which may intend to connect to the SBS providing highest QoS to that user. However, we discuss a scenario where individual QoS needs to be satisfied in addition to the network overall sum throughput in Section III.

Several papers consider the coordinated multi-point (CoMP) transmission [14], [15], which is shown to significantly improve the performance. However, such a strategy comes with a significant cost of training and feedback overhead for CSI. It is important to note that in future envisioned mmWave frequency bands, where channel statistics are rapidly varying, CSI feedback is not purposeful. Typical assignment strategies [16], [17] require channel conditions to be known and availability of CSI feedback. However, absence of CSI feedback necessitates an orthogonal resource block allocation to mitigate the inter-user and inter-SBS interference, and an online learning strategy to interact with the environment to solve the combinatiorial assignment problem with volatile user-SBS pairs.

Earlier works solve the user-SBS associations for multiple SBSs or user-channel association for single SBS without considering the user mobility or volatile nature of the network [5], [6], [13], [18]. It is worth noting that user mobility and volatility can quickly degrade the performance of the solution considered to be optimal for the static scenario. These issues motivate development of online learning strategies to cater mobility management. For instance, [7] solves the association for a single user in a dynamic environment, however, ignoring the volatility of associations. In contrast to [7], we consider mobility management for multiple users, where the mobility management system makes decisions about the users and their connection with the SBSs.

Recently, there has been a surge of interest in applying MAB algorithms, which are developed to maximize rewards in sequential decision-making problems in uncertain environments, to solve communication problems over dynamic environments [7], [19], [20]. Typical MAB settings involve a learner which selects an arm sequentially over rounds targeting to maximize the sum of rewards. The learner does not know the reward distributions beforehand, and in each round can only observe the reward from the selected arm (aka action). Due to this, it faces the typical dilemma of striking the balance between exploration and exploitation [21], [22]. Notable extensions of MAB are: i) Contextual MAB [23]–[25]: a side-information is observed by the learner at the beginning of a round, known as context, and the obtained reward is dependent on the selected arm and the observed context; ii) Combinatorial MAB [26]–[28]: instead of choosing a single arm in each round, a subset of base arms is selected, known as super arm, and the obtained reward is dependent on the elements (base arms) in the selected super arm in a semi-bandit feedback, i.e., only the rewards of the selected base arms are obtained; iii) Volatile MAB [29], [30]: a learner chooses an arm from a time varying arm set, and thus, the available arms may vary across rounds. In addition to works mentioned above, another line of work is multiplayer MAB (MPMAB), where multiple users select arms simultaneously in each round in order to maximize the sum of the rewards. For instance, Game of Thrones (GoT) in [31] is a fully distributed algorithm that solves the distributed assignment problem via collision and reward feedbacks. In [32], random exploration in GoT is replaced with channel orthogonalization in order to improve the reward estimation in exploration phase. Contrary to GoT, where collisions result in zero reward, authors in [33] consider the case in which colliding users receive non-zero rewards. The distributed algorithms discussed above provide static and non-contextual solutions to the maximal matching problem. In contrast, our proposed algorithm addresses the dynamic maximal matching problem by exploiting the contextual information.

The base station assignment (BSA) problem in the proposed setting shapes to a complex problem, due to rapidly varying channel conditions, absence of CSI feedback and dynamic users, and thus, is not solely solvable by combinatorial optimization techniques (e.g., Hungarian method [3]). Therefore, we propose an online contextual combinatorial volatile MAB algorithm with the fusion of contextual MAB, combinatorial MAB and volatile MAB to learn the best assignment strategy in an unknown stochastic environment with volatile user-SBS pairs via utilizing the orthogonal resource block allocation.

Our work is also related to a conference paper which presents theoretical work on contextual combinatorial volalite bandits [34]. This theoretical work models sequential decision-making in a combinatorial and volatile environment, and provides a learning algorithm with a sublinear regret bound without a concrete application. On the other hand, our paper's main objective is to design a computationally efficient, provably optimal learning algorithm for the user-SBS association problem. Therefore, we need to take into account the unique characteristics of this problem such as proving

TABLE I
COMPARISON OF THE PROPOSED FRAMEWORK WITH THE RELATED WORKS

Framework	Multiple Users	Mobility Management	Volatile user-SBS Associations	Contextual
Primal-Dual Distributed Algorithm [5]	✓	×	×	×
College Admission Game [6]	✓	×	×	×
Mobility Management [7]	×	✓	×	×
Iterative Improvement Algorithm [8]	✓	✓	×	×
MUSIC (Our work)	✓	✓	✓	✓

smoothness of the expected throughput in the context and handling different numbers of base associations in each round. While the learning algorithm in [34] works with an approximation oracle, in this paper we use a computationally efficient exact oracle. Our algorithm uses new confidence bounds derived from analysis of self-normalized martingales [35], which enables us to control the confidence level of our regret bounds. Our algorithm accumulates knowledge through contexts by performing an adaptive discretization of the context space [36]. Earlier works which used this technique considered MAB problems where the set of arms is large, but fixed (non-volatile), and a single arm is selected at each time step. In contrast, we are the first to apply adaptive discretization in a combinatorial learning problem where the arm set is dynamically changing over time. Thus, our discretization is affected by how contexts arrive over time. Moreover, different from previous works, our algorithm decides which set of arms to explore or exploit based on their total reward and total uncertainty. Last but not least, we provide extensive experiments and report various performance metrics that highlight applicability of the proposed algorithm to the user-SBS association problem. A comparison between our work and the related works is given in Table I.

II. PROBLEM FORMULATION

A. System Model

We consider a wireless network with M SBSs indexed by the set $\mathcal{M} := [M]$ and dynamically arriving/departing mobile users.[1] Each SBS $m \in \mathcal{M}$ can serve at most q_m users simultaneously. Furthermore, a user i can be served at most by a single SBS similar to earlier works [5]–[8]. While our formalism can be easily generalized to handle multiple SBSs association for a single user, this assumption is put since the multiple SBSs association for a single user requires more overhead to implement and may not be viable in practice [5].

The system operates in discrete rounds (time slots) indexed by $t \in [T]$, where T represents the time horizon. We assume that SBSs use orthogonal resource block allocation (RBs) based on time division multiple access (TDMA) or two-dimensional time-frequency RBs to avoid inter-SBS interference [8], [37]. We assume that the orthogonal RBs allow fixed airtime usage fraction for each SBS in the system and fixed modulation and/or coding scheme (MCS), and thus, the transmission rate r is fixed for all SBSs. Furthermore, SBS m adopts FDMA, i.e., the SBS's spectrum is divided into q_m channels, and thus, the number of users served by SBS m is upper bounded by q_m [13], [16], or it can allocate q_m two-dimensional time-frequency orthogonal resource blocks (RBs) [37]. Since each SBS can serve multiple users, we refer each of its partitioned slot (e.g., resource blocks (RBs) [37] or subcarriers [13]) as a channel, i.e., SBS channel (SBSC). The set of such channels for SBS m is denoted by $\mathcal{C}_m$ and its cardinality is restricted by the capacity q_m of SBS m, i.e., $\mathcal{C}_m := \{c_{m,1}, \ldots, c_{m,q_m}\}$.

[1]For an integer $M > 0$, $[M] = \{1, \ldots, M\}$.

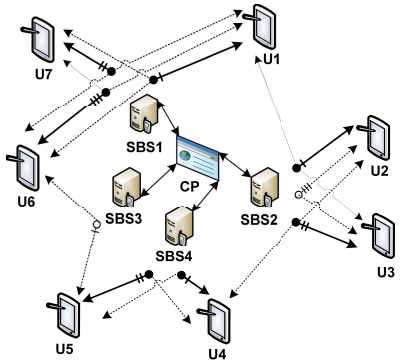

Fig. 1. There are 4 SBSs (9 SBSCs: 3 in SBS1, 3 in SBS2, 2 in SBS4, and 1 in SBS3) under a CP. For a particular round, 7 dynamic users are available in the area. In this example SBS1 serves U1, U6 and U7, SBS2 serves U2 and U3, and SBS4 serves U4 and U5. Solid lines represent the selected base associations, which constitute the current super association, dotted lines the non-selected base associations. The filled base of a connection represents a selected SBSC, whereas an empty base represents a non-selected SBSC. Some SBSCs, which are not part of the current super association, are not shown.

We consider a very general heterogeneous network, where a user may experience different channel gains, not only over SBSs but also over SBSCs [13], [37], [38]. Note that our proposed model includes as a special case a simplified scenario, where there is only a single SBS with numerous SBSCs. We represent the set of SBSCs by $\overline{\mathcal{C}} := \cup_{m \in \mathcal{M}} \mathcal{C}_m$. Since the set of users in the system changes dynamically over time, we represent the set of available users in round t by $\mathcal{N}_t$. Similarly, the set of available user-SBSC pairs in round t is represented by $\mathcal{A}_t := \{(c, n) : c \in \overline{\mathcal{C}}, n \in \mathcal{N}_t\}$. We refer to the set $\mathcal{A}_t$ as the *base associations set* and each element in this set as a *base association*.

At the beginning of each round t, available users are associated with SBSCs by a central processor (CP). Then, at the end of round t, the reward (throughput) associated with the chosen user-SBSC association is obtained. Fig. 1 represents an example of a user-SBSC association in a particular round. We assume that the CP has access to contextual information for each $a \in \mathcal{A}_t$, which is represented by $x_{t,a}$. For all $a \in \mathcal{A}_t$ and $t > 0$, the context $x_{t,a}$ lies in a predefined context space $(\mathcal{X}, \bar{d})$. Here, $\mathcal{X}$ represents the context set and $\bar{d}$ is the distance metric, where $\bar{d}(x, x')$ represents the distance between contexts x and x', $\forall x, x' \in \mathcal{X}$.

For each user-SBSC association, we assume that the transmit power is fixed as p and the transmission rate is fixed as r. We also assume that channel gains $h_a, \forall a \in \mathcal{A}_t, \forall t > 0$, are i.i.d. random variables, which is a widespread assumption in the channel allocation literature [13], [38], [39]. We do not assume channel gains to be strictly Rayleigh fading, as they can belong to a much broader class of distributions such as Nakagami-m, Rician and others [13]. We assume that the heterogeneity in channel gains is captured by the transmission frequency, and the user-SBS distance as discussed in Section II-D. Therefore, for each user-SBSC pair, the context is defined as $x := [f/f_{\max}, d/d_{\max}]$, where $f > 1$GHz is the transmission frequency (in GHz), $d > 1$m is the distance (in meters) between the user and the corresponding SBS (to whom the SBSC belongs), $f_{\max}$ is the normalization factor for frequency, and $d_{\max}$ is the normalization factor for distance to ensure the fact that all contexts lie in the unit square. Thus, the context space $\mathcal{X}$ is in $[0,1]^2$. The distance metric is defined as $\bar{d}(x,x') = ||x - x'||_1$, also known as Manhattan distance or taxicab norm.[2]

B. Objective, Rewards, Super Associations

The objective of the CP is to maximize the long-term sum throughput of all the users via utilizing the contextual information of volatile user-SBSC pairs. To reduce the computational overhead, the CP in the proposed system is the central entity only to perform the assignment of SBSCs to users and SBSs only share their rewards with CP, whereas decoding of the data is performed by the corresponding SBS, distributedly. The feedback of the reward is also limited; the reward feedback to the base stations and the CP is considered to be only a one-bit feedback, which provides the information about success or failure of the transmitted data block.

The (random) throughput of user-SBSC pair $a \in \mathcal{A}_t$ in round t is denoted by $g(x_{t,a}) \in \{0, r\}$. Similarly, we define the expected throughput of user-SBSC pair a with context x_a as

$$\mu(x_a) := \mathbb{E}[g(x_a)] = r\theta(x_a) \tag{1}$$

where $\theta(x_a)$ represents the transmission success probability under context x_a. Since the rate is fixed, we can normalize the expected throughput by diving it with r, which is equivalent to taking $r = 1$ in the remainder of the paper. Note that the throughput can be written as $g(x_{t,a}) = \mu(x_{t,a}) + \eta_t$, where η_t is a 1-sub-Gaussian random variable that represents the noise.[3]

We define $K_t := \min\{|\mathcal{N}_t|, \overline{C}\}$ as the current *occupancy* of the system in round t, to represent the case when users are lower in number than the SBSCs, and vice-versa. The maximum occupancy is defined as $K := \sum_{m \in \mathcal{M}} q_m$, and we have $K_t \leq K, \forall t \leq T$. A *super association* is a subset of the base associations set ($S \subseteq \mathcal{A}_t$), which satisfies the following conditions:

i) A user or an SBSC can appear only once.

ii) The size of the subset is equal to the occupancy of the system in round t i.e., $|S| = K_t$.

Thus any $S \subseteq \mathcal{A}_t$ satisfies the following:

$$S = \big\{(c_i, n_i), i \in [K_t] :\ c_i \in \overline{\mathcal{C}}, n_i \in \mathcal{N}_t \mid (c_i \neq c_j)\ \&\ (n_i \neq n_j), \forall (i \neq j) \in [K_t]\big\}.$$

The set of such super associations in round t is referred to as *super association set* and is denoted by $\mathcal{S}_t$. Its cardinality is denoted by $P(\overline{C}, N_t)$. For example, if we have only 6 SBSs (e.g., 16 SBSCs) and 6 users, the cardinality of super association set is $5,765,760$ super associations. Due to the huge number of super associations, it is impossible to learn the optimal super associations by using the traditional MAB formulation in a reasonable amount of rounds. On the other hand, our algorithm will exploit the fact that base associations which create super associations are very few in number, i.e., $(16 \times 6) = 96$ base associations in the considered example. Thus, exploring these base associations is sufficient to reach to the optimal super association, since every super association is a permutation of the available base associations. The super association selected from $\mathcal{S}_t$ by the CP in round t is denoted by S_t.

Our association problem can also be viewed as a dynamic maximal matching problem in the bipartite graph formed by the users and SBSCs with unknown weights. In this graph, nodes correspond to users and SBSCs, edges correspond to user-SBSC pairs, and edge weights correspond to throughputs. Unlike static maximal matching, weights associated with the edges vary from one round to another. Thus, the maximal matching (the optimal super association) varies over time. Since the expected throughput of a user-SBSC pair depends only on its context, we can also represent each super association S by the contexts of its base associations. Therefore, super association S in round t is also defined by the context tuple $\mathbf{x}_{t,S} := (x_{t,a})_{a \in S}$. For this super association, the corresponding random throughput and expected throughput vectors are given as $g(\mathbf{x}_{t,S}) := (g(x_{t,a}))_{a \in S}$ and $\mu(\mathbf{x}_{t,S}) := (\mu(x_{t,a}))_{a \in S}$. The corresponding throughput received from selecting this super association is the summation of throughputs of the base associations and is denoted by $u(g(\mathbf{x}_{t,S}))$, i.e., $u(g(\mathbf{x}_{t,S})) = \sum_{a \in S} g(x_{t,a})$. Due to linearity of expectation, we also have $\mathbb{E}[u(g(\mathbf{x}_{t,S}))] = \sum_{a \in S} \mu(x_{t,a}) = u(\mu(\mathbf{x}_{t,S}))$. Finally, we denote the set of available contexts in round t by $\mathcal{X}_t := \{x_{t,a}\}_{a \in \mathcal{A}_t}$ and the vector of expected throughputs of the available base associations by $\mu_t := [\mu(x_{t,a})]_{a \in \mathcal{A}_t}$.

C. Cumulative Expected Throughput and the Regret

The cumulative expected throughput of the selected super associations by round T is defined as $\tilde{\mathbf{W}}(T) := \sum_{t=1}^{T} u(\mu(\mathbf{x}_{t,S_t}))$. The optimal super association in round t is given as $S_t^* := \operatorname{argmax}_{S \in \mathcal{S}_t} u(\mu(\mathbf{x}_{t,S}))$ and its expected throughput is denoted by $opt(\mu_t) := u(\mu(\mathbf{x}_{t,S_t^*}))$. For simplicity, we assume that S_t^* is unique. S_t^* can be computed

[2]The regret analysis in Section IV holds for any finite dimensional $\mathcal{X}$ and any distance metric defined over $\mathcal{X}$.

[3]Let X be a Bernoulli random variable with parameter θ. It holds that $X = \theta + \eta$, where η is a random variable that takes values $-\theta$ or $1-\theta$. Hoeffding's lemma states that every random variable bounded in the interval $[a,b]$ is sub-Gaussian with parameter $(b-a)/2$. It follows that η is sub-Gaussian with parameter $1/2$, and hence, also with parameter 1, since the condition of sub-Gaussianity is still satisfied, that is $\ln \mathbb{E}[e^{\lambda(\eta - \mathbb{E}[\eta])}] \leq \frac{\lambda^2}{4}$, for all $\lambda \in \mathbb{R}$, implying that $\ln \mathbb{E}[e^{\lambda(\eta - \mathbb{E}[\eta])}] \leq \frac{\lambda^2}{2}$, for all $\lambda \in \mathbb{R}$.

in polynomial time when μ is perfectly known by Hungarian method or its variants [3], [4]. Thus, we assume that the CP has access to an exact oracle based on Munkres' variant of Hungarian method, which, when given as input μ_t returns an optimal solution. We let $x^*_{t,k}$ represent the context of the kth base association in S^*_t.

Since the CP does not know μ_t in our case, it gives an $|\mathcal{A}_t|$-dimensional parameter vector ϑ_t that consists of indices of the available base associations as input to the exact oracle to get $S_t = \text{Oracle}(\vartheta_t)$.[4] We let $\tilde{x}_{t,k}$ represent the context of the kth selected base association by the Oracle in round t. Note that S_t is an optimal solution under ϑ_t but no necessarily under μ_t. To characterize the loss of the CP due to not knowing the expected throughputs of user-SBSC pairs for all possible contexts, we define the regret by round T as

$$R(T) := \sum_{t=1}^{T} opt(\mu_t) - \sum_{t=1}^{T} u(\mu(\mathbf{x}_{t,S_t})) \ .$$

Thus, maximizing the sum throughput is equivalent to minimizing the growth rate of the regret.

D. Form and Smoothness of the Expected Throughput

Minimizing the regret without knowledge of μ is a non-trivial task since the CP needs to balance exploration and exploitation of the available user-SBSC pairs. In order to utilize the information gained from the past user-SBSC associations, the CP needs to exploit the similarities between past and present contexts. Next, we describe in detail how the expected throughput depends on the context. This requires us to formally introduce the channel model. In particular, we consider a mmWave channel model with three link states:

- Line-of-sight (LOS): Occurs when there is no blockage between the SBS and the user.
- Non-line-of-sight (NLOS): Occurs when the direct LOS between the SBS and the user is blocked.
- Outage: Occurs when there is no link between the SBS and the user, i.e., the path loss is very high.

Let a be the base association that represents the link between user n and SBSC c_m of SBS $m \in \mathcal{M}$. The signal-to-interference-plus-noise ratio (SINR) for a is given as $\gamma_a = \frac{p h_a}{\sigma_a^2 + I_a}$, where p is the transmit power, h_a is the channel gain, I_a is the interference from other SBSs and $\sigma_a^2 = \sigma^2$ is the variance of the additive noise. We assume that the interference between SBSs is well managed (e.g., dedicated bands, orthogonal resource blocks [40], [41]), and thus, we set $I_a = 0$.

The channel gain over base association a in LOS state is given as $h_a^{LOS} = c_{LOS}\big(\frac{1}{d_a}\big)^{\eta_{LOS}} Y_{LOS}$, where c_{LOS} is the average multiplicative gain at the reference distance d_0 ($d_0 = 1$ meter), η_{LOS} is the pathloss exponent, d_a is the distance in meters between the user and the SBS for base association a, and $Y_{LOS} \geq 0$ is the random variable that reflects the fading distribution [42], [43]. We have $c_{LOS} = d_0^{\eta_{LOS}} \frac{1}{FSPL(f_a, d_0)}$, where $FSPL(f_a, d_0)$ is the path loss in free space, given as $FSPL(f_a, d_0) = \big(\frac{4\pi d_0 \ f_a \times 10^9}{3\times 10^8}\big)^2 = \big(\frac{4\pi d_0 \ f_a}{0.3}\big)^2$ (antennas are assumed to have unity gain), where f_a is the transmission frequency in GHz. Thus, $\gamma_a^{LOS} = G_{LOS} \frac{1}{f_a^2 \ d_a^{\eta_{LOS}}} Y_{LOS}, d_a \geq d_0$ where $G_{LOS} := (p(0.3)^2 \ d_0^{\eta_{LOS}-2})/(16\pi^2\sigma^2)$ is a constant. In logarithmic scale, we have [44], [45]

$$h_a^{LOS}[dB] = 20\log_{10}\big(\frac{0.3}{4\pi d_0}\big) - 20\log_{10}(f_a) - 10\eta_{LOS}\log_{10}(\frac{d_a}{d_0}) - \xi_{LOS} \quad (2)$$

where $-\xi_{LOS} = 10\log_{10}(Y_{LOS})$ and $Y_{LOS} = 10^{\frac{-\xi_{LOS}}{10}}$.

Similarly, the channel gain over base association a in NLOS state in logarithmic scale is given as

$$h_a^{NLOS}[dB] = 20\log_{10}\big(\frac{0.3}{4\pi d_0}\big) - 20\log_{10}(f_a) - 10\eta_{NLOS}\log_{10}(\frac{d_a}{d_0}) - \xi_{NLOS} \quad (3)$$

where $-\xi_{NLOS} = 10\log_{10}(Y_{NLOS})$. Following the same analysis for LOS, we have $\gamma_a^{NLOS} = G_{NLOS}\frac{1}{f_a^2 \ d_a^{\eta_{NLOS}}} Y_{NLOS}, d_a \geq d_0$ for NLOS, where $G_{NLOS} := (p(0.3)^2 \ d_0^{\eta_{NLOS}-2})/(16\pi^2\sigma^2)$. The channel gain over base association a in outage state is given as $h_a^O = 0$, which in turn gives $\gamma_a^O = 0$. Thus, heterogeneity in channel gains is sufficiently captured by parameters f_a and d_a. We consider the context as tuple given by $[f/f_{\max}, d/d_{\max}]$. For other models like floating-intercept (FI)/ABG, our algorithm will work and our results will hold as long as the expected throughput varies smoothly with the context, since frequency and distance sufficiently cover the studied models [44]–[47]. For instance, ABG model in [45], [46], which is given as $h_a^{ABG}(f_a, d_a)[dB] = 10\eta\log_{10}(d_a) + \beta + 10\gamma\log_{10}(f_a) + X_{ABG}$ is dependent on both the transmission frequency and SBS-user distance which constitute the context in our setting. Similarly, modified intercept model in [44] and path loss model in [47], given by $h_a(d_a)[dB] = \alpha + 10\beta\log_{10}(d_a) + \xi$ provide the fact that α and β are curve fitting parameters characterized by considered context components.

The probabilities of LOS, NLOS, and outage are given as [48], [49]

$$\begin{aligned} p_O(d_a) &= \max(0, 1 - \zeta_O e^{-\delta_O d_a}); \\ p_{NLOS}(d_a) &= (1 - p_O(d_a))(1 - \zeta_{LOS} e^{-\delta_{LOS} d_a}); \\ p_{LOS}(d_a) &= (1 - p_O(d_a))\zeta_{LOS} e^{-\delta_{LOS} d_a}; \end{aligned} \quad (4)$$

The parameters ζ_O, ζ_{LOS}, δ_O, and δ_{LOS} are dependent on the transmission frequency and propagation scenario (see [48, Table I]). Note that none of these are required to be known by the CP. We only use them to characterize the smoothness of the expected throughput.

Using above given probabilities, we can write

$$\begin{aligned} \gamma_a = \ & \mathbf{1}[U < p_{LOS}(d_a)]\gamma_a^{LOS} \\ & + \mathbf{1}[p_{LOS}(d_a) \leq U < (p_{LOS}(d_a) + p_{NLOS}(d_a))]\gamma_a^{NLOS} \\ & + \mathbf{1}[(p_{LOS}(d_a) + p_{NLOS}(d_a)) \leq U \leq 1]\gamma_a^O \end{aligned} \quad (5)$$

where $\mathbf{1}$ is the indicator function, and $U \in [0, 1]$ is a uniform random variable. The next assumption connects the

[4]Indices are chosen as upper confidence bounds on the expected throughputs. Their exact form is given in Section III.

transmission success probability with the SINR. With an abuse of notation we represent the transmission success probability given SINR γ as $\theta(\gamma)$.[5]

Assumption 1: The success probability $\theta(\gamma)$ as a function of γ is well-approximated by the exponential function, and is given as $\theta(\gamma) = (1 - \exp(-\gamma))$ [50], [51].[6]

Letting $f_U(u) = 1, 0 \leq u \leq 1$, and using (5) and Assumption 1, we obtain

$$\begin{aligned}\mu(x_a) &= r\mathbb{E}[\theta(\gamma_a)|x_a] \\ &= r\big(\mathbb{E}[\theta(\gamma_a)|\gamma_a = \gamma_a^{LOS}]\mathbb{P}(\gamma_a = \gamma_a^{LOS}) \\ &\quad + \mathbb{E}[\theta(\gamma_a)|\gamma_a = \gamma_a^{NLOS}]\mathbb{P}(\gamma_a = \gamma_a^{NLOS})\big) \\ &= r\int_0^{p_{LOS}(d_a)} \mathbb{E}_{Y_{LOS}}[(1 - \exp(\gamma_a^{LOS}))]du \\ &\quad + r\int_{p_{LOS}(d_a)}^{(p_{LOS}(d_a)+p_{NLOS}(d_a))} \\ &\quad \times \mathbb{E}_{Y_{NLOS}}[(1 - \exp(\gamma_a^{NLOS}))]du \\ &= p_{LOS}(d_a)r\mathbb{E}_{Y_{LOS}}[(1 - \exp(\gamma_a^{LOS}))] \\ &\quad + p_{NLOS}(d_a)r\mathbb{E}_{Y_{NLOS}}[(1 - \exp(\gamma_a^{NLOS}))]. \end{aligned} \tag{6}$$

Hence, the expected reward for a context $x_a = [\frac{f}{f_{\max}}, \frac{d}{d_{\max}}]$ can be written as

$$\begin{aligned}&\mu(x_a) \\ &= p_{LOS}(d_a)\int_y r\big(1 - \exp(-G_{LOS}w_{LOS}(x_a)y)\big)f_{Y_{LOS}}(y)dy \\ &\quad + p_{NLOS}(d_a)\int_y r\big(1 - \exp(-G_{NLOS}w_{NLOS}(x_a)y)\big) \\ &\quad \times f_{Y_{NLOS}}(y)dy \end{aligned} \tag{7}$$

where $w_{LOS}(x_a) = \frac{1}{f_a^2}\frac{1}{d_a^{\eta_{LOS}}}$ and $w_{NLOS}(x_a) = \frac{1}{f_a^2}\frac{1}{d_a^{\eta_{NLOS}}}$. By using $\mu_{LOS}(x_a) = r\int_y \big(1 - \exp(-G_{LOS}\frac{1}{f_a^2}\frac{1}{d_a^{\eta_{LOS}}}y)\big)f_{Y_{LOS}}(y)dy$ and $\mu_{NLOS}(x_a) = r\int_y \big(1 - \exp(-G_{NLOS}\frac{1}{f_a^2}\frac{1}{d_a^{\eta_{NLOS}}}y)\big)f_{Y_{NLOS}}(y)dy$ we can rewrite (7) as,

$$\mu(x_a) = p_{LOS}(d_a)\mu_{LOS}(x_a) + p_{NLOS}(d_a)\mu_{NLOS}(x_a). \tag{8}$$

The main result of this section is given in the following lemma whose proof can be found in Appendix B.

Lemma 1: Let $\bar{d}(x, x') = ||x - x'||_1$. Then, for all $x, x' \in \mathcal{X}$, we have $|\mu(x) - \mu(x')| \leq L\bar{d}(x, x')$, for some Lipschitz constant $L > 0$.

Lipschitz continuity is a mild form of continuity that manifests itself in many practical applications. For our application, it merely implies that the expected throughput of a base association smoothly varies as a function of the context. Indeed, this is what we proved in Lemma 1. Since the distance between the user and the SBS is a smooth function of the position of the user, one can also show that the expected throughput varies smoothly with the position of the user.

Remark 1: The proposed algorithm and its theoretical analysis are applicable to any finite dimensional context space, and thus, when distance dependent blockage probabilities are unknown, additional dimensions to the context can be added to solve the stated problem. For instance, we provide experimental results in Section V-D, where angle between the user and the SBS is taken as the additional 3rd dimension to the context. This context-trio helps to learn not only frequency and distance dependent throughput, but also position dependent throughput.

[5]Note that we have $\theta(x_a) = \mathbb{E}[\theta(\gamma)|x_a]$.

[6]While we choose $(1 - \exp(-\gamma))$ as an approximation to the probability of correct symbol reception in Lemma 1, a similar smoothness result can also be shown for $\theta(\gamma) = (1 - \exp(-\gamma))^s$, which is a good approximation when each data packet contains s symbols.

E. Properties of the Context Space

Since $\mathcal{X}$ is in $[0, 1]^2$, it satisfies some regularity conditions. Our algorithm adaptively discretizes the context space in order to navigate its exploration based on how contexts have arrived in the past. Performing adaptive discretization requires that the context space is *well-behaved*.

Definition 1: (Well-behaved metric space [52]) A compact metric space $(\mathcal{X}, \bar{d})$ is said to be well-behaved if there exists a sequence of subsets $(\mathcal{X}^h)_{h\geq 0}$ of $\mathcal{X}$ satisfying the following properties:

1) *There exists $Z \in \mathbb{N}$, such that each subset $\mathcal{X}^h$ has Z^h elements, i.e. $\mathcal{X}^h = \{x^{(h,i)},\ 1 \leq i \leq Z^h\}$ and each element $x^{(h,i)}$ is associated with a cell $X^{(h,i)} = \{x \in \mathcal{X} :\ \bar{d}(x, x^{(h,i)}) \leq \bar{d}(x, x^{(h,j)}),\ \forall j \neq i\}$.*
2) *For all $h \geq 0$ and $1 \leq i \leq Z^h$, we have: $X^{(h,i)} = \cup_{j=Z(i-1)+1}^{Zi} X^{(h+1,j)}$. The nodes (quantized contexts) $x^{(h+1,j)}$ for $Z(i-1)+1 \leq j \leq Zi$ are called the children of $x^{(h,i)}$, which in turn is referred to as the parent.*
3) *We assume that the cells have geometrically decaying radii, i.e., there exists $0 < \rho < 1$ and $0 < v_2 \leq 1 \leq v_1$ such that we have $O(x^{(h,i)}, v_2\rho^h/2) \subseteq X^{(h,i)} \subseteq O(x^{(h,i)}, v_1\rho^h/2)$, where $O(i, j)$ denotes the closed ball with center at i and radius j. Note that we have $v_2\rho^h \leq \text{diam}(X^{(h,i)}) \leq v_1\rho^h$, where $\text{diam}(X^{(h,i)}) := \sup_{x,y\in X^{(h,i)}} \bar{d}(x, y)$.*

The first property implies that for every $h \geq 0$ the cells $X^{(h,i)}, 1 \leq i \leq Z^h$ partition $\mathcal{X}$. The second property intuitively means that as h grows, we get a more refined sequence of partitions. The third property implies that the quantized contexts $x^{(h,i)}$ are evenly spread out in the space. As previously indicated in [36], [52] for general Euclidean norms, our context space $(\mathcal{X}, \bar{d})$ is well behaved.

An important expression that will appear in our theoretical analysis is the approximate-optimality dimension [34] associated with $(\mathcal{X}, \bar{d})$, μ and the sequence of contexts. The approximate-optimality dimension can be viewed as the dimension of the space where the approximately-optimal contexts related to μ for the given sequence of contexts reside. Thus, it captures both the structure of $(\mathcal{X}, \bar{d})$ and μ, and the specific context sequences encountered during the learning process. Implicitly taking into account the dependence on these quantities, we represent the approximate-optimality dimension by $\bar{D}$ (details are given in Appendix A). It has the nice property that $\bar{D} \leq D$, where D is the usual dimension. Moreover, it can even be much smaller than D in favorable cases.

III. Algorithm

Our algorithm is called *Multi-User Small base stations assocIation via adaptive Contextual combinatorial strategy* (MUSIC). Its pseudocode is given in Algorithm 1. At the beginning of round t, MUSIC observes available base associations (since some of base associations are unavailable due to dynamic users) and their corresponding contexts. Each context is associated with an active quantized context (leaf node) $x^{(h,i)}$ given in Definition 1. The set of active quantized contexts in round t, denoted by $\mathcal{L}_t$, forms a partition of $\mathcal{X}$. We denote the parent of quantized context $x^{(h,i)}$ by $p(x^{(h,i)})$. Let $\mathcal{O}_t$ denote the set of available active quantized contexts, whose regions contain the available contexts for the available base associations. For each active quantized context, we maintain an index which is an *upper confidence bound* (UCB) on the maximum expected throughput of base associations which have contexts in the region associated with the quantized context.

Before defining the index of a given quantized context $x^{(h,i)}$, we first define the auxiliary terms which comprise it. The term

$$b_t(x^{(h,i)}) := \min\{\hat{\mu}_{t-1}(x^{(h,i)}) + c_{t-1}(x^{(h,i)}), \hat{\mu}_{t-1}(p(x^{(h,i)})) + c_{t-1}(p(x^{(h,i)})) + Lv_1\rho^{h-1}\}$$

is a high probability upper bound on $\mu(x^{(h,i)})$, where

$$c_t(x^{(h,i)}) = \sqrt{\frac{(1+B_t(x^{(h,i)}))}{(B_t(x^{(h,i)}))^2}\left(1+2\log\left(\frac{KT(1+B_t(x^{(h,i)}))^{1/2}}{\delta}\right)\right)} \tag{9}$$

is the confidence radius, tailored to give high probability upper bounds on μ. Here $B_t(x^{(h,i)})$ is the number of times a base association with a context from the cell $X^{(h,i)}$ was selected by the CP, formally defined as $B_t(x^{(h,i)}) := \sum_{t'=1}^{t}\sum_{k=1}^{K_t}\mathbb{I}\{(H_{t',k}, I_{t',k}) = (h,i)\}$ where we denote by $(H_{t',k}, I_{t',k})$ the active quantized context associated with the cell containing the context of the kth selected base association in round t', and $\delta \in (0,1)$. We define the total reward accumulated by the CP until round t from selecting associations with contexts associated with the quantized context $x^{(h,i)}$ as follows: $v_t(x^{(h,i)}) := \sum_{t'=1}^{t}\sum_{k=1}^{K_t} g(\tilde{x}_{t',k})\mathbb{I}\{(H_{t',k}, I_{t',k}) = (h,i)\}$. Consequently, the empirical mean is defined as

$$\hat{\mu}_t(x^{(h,i)}) := \begin{cases} v_t(x^{(h,i)})/B_t(x^{(h,i)}) & \text{for } B_t(x^{(h,i)}) > 0 \\ 0 & \text{otherwise} \end{cases}$$

The constants v_1 and ρ are parameters associated with the metric space that are given as input to the CP as specified in Definition 1.

Definition 2: The index of quantized context $x^{(h,i)}$ is defined as $\overline{\mu}_t(x^{(h,i)}) := b_t(x^{(h,i)}) + Lv_1\rho^h$. Moreover, let $\phi_{t,a} = x_{\tilde{h}_{t,a},\tilde{i}_{t,a}}$ be the active quantized context associated with the cell that contains $x_{t,a}$. The index of a base association $a \in \mathcal{A}_t$ is defined as $\overline{\mu}_t(x_{t,a}) := \overline{\mu}_t(\phi_{t,a}) + (Zv_1/v_2)Lv_1\rho^{\tilde{h}_{t,a}}$, where the second term guarantees (with high probability) that $\overline{\mu}_t(x_{t,a})$ upper bounds $\mu(x_{t,a})$ and Z is a natural number given in Definition 1.

Algorithm 1 MUSIC

Require: $\mathcal{X}$, $(\mathcal{X}^h)_{h\geq 0}$, v_1, v_2, ρ, Z, L, K, T, δ.
Initialize: $B_0(x^{(h,i)}) = 0$, $\hat{\mu}_0(x^{(h,i)}) = 0$, $\forall\ x^{(h,i)} \in \mathcal{X}$; $\mathcal{X}^0 = \mathcal{X}$, $\mathcal{L}_1 = \{x^{(0,1)}\}$.
for $t = 1, 2, \ldots, T$ **do**:
  Observe base associations in $\mathcal{A}_t$ and their contexts $\mathcal{X}_t$.
  Identify available active quantized contexts $\mathcal{O}_t \subseteq \mathcal{L}_t$.
  $S_t \leftarrow$Computing the maximal matching using Munkres' Algorithm $((\overline{\mu}_t(x_{t,a}))_{a\in\mathcal{A}_t})$ [4].
  Inform users and SBS about their associations.
  Obtain throughput feedback $(g(x_{t,a}))_{a\in S_t}$ from the users.
  Identify the set of selected quantized context $\mathcal{P}_t$.
  for $x^{(h,i)} \in \mathcal{P}_t$ **do**:
    Update $\hat{\mu}_t(x^{(h,i)})$ as in (10), and $B_t(x^{(h,i)})$ as in (11).
    if $c_t(x^{(h,i)}) \leq Lv_1\rho^h$ **then**:
      $\mathcal{L}_{t+1} \leftarrow \mathcal{L}_t \cup \{x^{(h+1,j)} : Z(i-1)+1 \leq j \leq Zi\} \setminus \{x^{(h,i)}\}$.
    end if
  end for
end for

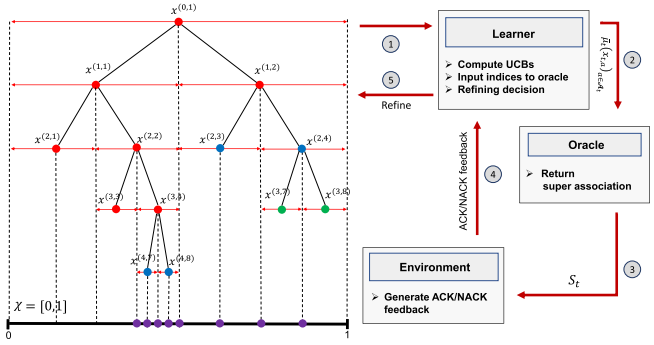

Fig. 2. Graphical representation of MUSIC algorithm in the case when $\mathcal{X} = [0,1]$ for a given round $t \geq 1$. The circled numbers correspond to the sequence of actions that MUSIC executes in round t. The red dots represent the quantized contexts of the tree, the blue dots represent the active quantized contexts in round t, the purple dots represent the available contexts $\mathcal{X}_t$ in round t and the green dots represent the quantized contexts created by expanding $x^{(2,4)}$ at the end of round t.

Note that all base associations with contexts associated to the same quantized context have equal indices. A base association $x_{t,a}$ will have a high index when the active quantized context $x^{(h,i)}$ that it belongs to (i) does not have enough data samples to guarantee accurate estimation of the throughput, i.e., $c_t(x^{(h,i)})$ is large; or (ii) contains a wide range of contexts, i.e., h is small; or (iii) $c_t(x^{(h,i)})$ is small, h is large and $\hat{\mu}_t(x^{(h,i)})$ is large. If the first two conditions hold, then, $x_{t,a}$ is likely to be selected for exploration purposes. If the last condition holds, then, $x_{t,a}$ is likely to be selected for exploitation. We will see in the next section that this choice of index balances exploration and exploitation in the optimal way.

After the indices of the available base associations, i.e., $\{\overline{\mu}_t(x_{t,a})\}_{a\in\mathcal{A}_t}$ are computed, they are given as input to the Munkres' oracle in round t to obtain the super association $S_t \in \mathcal{S}_t$ that will be selected in round t. The Munkres' oracle is James Munkres' variant [4] of the Hungarian assignment

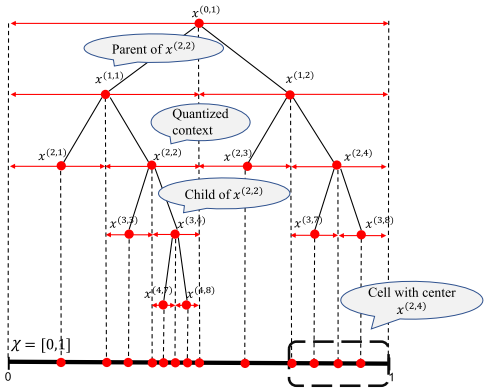

Fig. 3. Graphical representation of parent child relationship in context partition in the case when $\mathcal{X} = [0, 1]$.

problem [3]. Munkres' Assignment Algorithm achieves a low order polynomial run time, i.e., worst-case $O(n^3)$, where $n := \max\{N_t, \overline{C}\}$ is the dimension of the matrix given to Munkres' Assignment algorithm.

When S_t is selected by MUSIC, users and SBSs are notified of their associations in S_t. Afterwards, packet transmissions take place and each user records its throughput $g(x_{t,a})$ that corresponds to its association $a \in S_t$, and then, reports it to the CP at the end of round t.

In the learning step, MUSIC identifies the active quantized contexts that are selected, denoted by $\mathcal{P}_t$, and updates their statistics. For each $x^{(h,i)} \in \mathcal{P}_t$:

$$\hat{\mu}_t(x^{(h,i)}) = \frac{B_{t-1}(x^{(h,i)})\hat{\mu}_{t-1}(x^{(h,i)}) + rew_t(x^{(h,i)})}{B_{t-1}(x^{(h,i)}) + num_t(x^{(h,i)})} \quad (10)$$

$$B_t(x^{(h,i)}) = B_{t-1}(x^{(h,i)}) + num_t(x^{(h,i)}) \ , \quad (11)$$

where $rew_t(x^{(h,i)}) = \sum_{k=1}^{K_t} g(\tilde{x}_{t,k})\mathbb{I}((H_{t,k}, I_{t,k}) = (h,i))$ and $num_t(x^{(h,i)}) = \sum_{k=1}^{K_t} \mathbb{I}((H_{t,k}, I_{t,k}) = (h,i))$. Statistics of the other active quantized contexts do not change. Subsequently, for each quantized context $x^{(h,i)} \in \mathcal{P}_t$, MUSIC decides to expand it into Z children quantized contexts if $c_t(x^{(h,i)}) \leq Lv_1\rho^h$. Thus, the quantized context $x^{(h,i)}$ is expanded into Z children quantized contexts[7] $\{x^{(h+1,j)} : Z(i-1)+1 \leq j \leq Zi\}$ which are added to the set of active quantized contexts, whereas $x^{(h,i)}$ is removed from it. This allows MUSIC to tradeoff exploring base associations with new contexts and exploiting base associations with well-known contexts.

Finally, we note that optimistic algorithms that use UCB indices are widely used in learning problems in wireless communications. We use UCB-based algorithms since the uncertainty about expected throughputs of base associations is encoded explicitly in the exploration bonus. We make use of these bonuses in deciding when to refine the context partition. Thus, adaptive discretization technique is naturally tied with UCB indices.

Remark 2: The throughput maximization problem can be extended to handle the case where individual user QoS needs to be satisfied. In MAB literature, fair bandits [53]–[55] are able to solve individual constraints along with global goal maximization. The idea of volatility is used in fair bandits, i.e., when an individual arm is estimated to have lower individual expected reward than a threshold, this arm is considered unavailable for the current time slot. Therefore, our algorithm which is already able to cater to the volatility can easily handle this case. When volatile user-SBSCs are provided to the algorithm, additional constraint-dependent comparison is added to select only those user-SBSCs which are able to satisfy the individual QoS constraint with high probability. Mathematically, we compare the upper bound on estimated reward $\overline{\mu}_t(x^{(h,i)})$, for a user-SBSC pair whose context lie in $x^{(h,i)}$, with the provided QoS constraint ζ, and only those user-SBSCs are considered available who satisfy $\overline{\mu}_t(x^{(h,i)}) > \zeta$ for a given time slot t, which results in QoS being satisfied with high probability.

Remark 3: On the one hand, coherence time for mmWave is in the order of few ms, on the other hand, multi-Gigabit transmission can be achieved using mmWave, and therefore, multiple Giga bits per second transmission in mmWave allows a reasonable length of data packet to easily match the coherence time of mmWave. In contrast to known channel distribution and CSI feedback, which solves the optimization for each channel state, the contextual MAB handles unknown channel distribution and predicts the action which maximizes the reward for a given context in expectation. In contrast to slowly-varying wireless channels, where CSI feedback is useful, it may not be meaningful in mmWave where previously acquired CSI becomes outdated quickly. In addition, location information (contextual information which can easily be obtained by global positioning system) is effectively used instead of acquiring CSI which is resource intensive and computationally expensive.

IV. Theoretical Performance Analysis

A. Main Results

In this section, we show that the regret of MUSIC grows sublinearly over time by proving in Theorem 1, a regret bound of order $\tilde{O}(T^{(\bar{D}+1)/(\bar{D}+2)+\epsilon})$ for any $\epsilon > 0$, where $\bar{D}$ represents the approximate-optimality dimension. This implies that time-averaged regret $R(T)/T$ approaches to zero as T increases, which means that the average learning loss of MUSIC diminishes over time. Then, we show in Corollary 1 that when the contexts arrive from a finite subset of the unit square, $\bar{D} = 0$, so that the regret becomes $\tilde{O}(T^{1/2+\epsilon})$. For instance, this will be the case when the transmission frequency and distance between the user and the SBS take only finitely many different values. In practice, it holds when the frequency spectrum is divided into a predetermined set of bands and d can only be estimated up to a certain resolution [8], [13], [37]. This turns our problem into the volatile version of the finite-armed bandit for which the minimax lower bound is $\Omega(T^{1/2})$ [56]. Proofs of the main results are given in the Appendix.

Theorem 1: Fix $T > e/v_3$, where $v_3 := \frac{K\sqrt{e}}{Lv_1\delta}$. Given the parameters of the problem $Z \in \mathbb{N}$, $K \in \mathbb{N}$, $B > 0$, $L > 0$ and $0 < v_2 \leq 1 \leq v_1$, define $\bar{D} := D^{\mu}(f)$ (see Appendix A) and $f(r) = cr$, where $c = K(6Zv_1/v_2 + 2)(v_1/v_2)L$. Then, for any $D_1 > \bar{D}$ and $\delta \in (0,1)$, there exists $Q = Q(\mathcal{X}, \mu, c) > 0$

[7]See Definition 1.

(independent of T), for which the regret incurred by MUSIC is upper bounded with probability at least $1-\delta$ as follows:

$$\begin{aligned} R(T) \leq\ & C_0(D_1)\cdot T^{\frac{D_1-1}{D_1+2}}\cdot(\log(Tv_3))^{\frac{D_1+2}{D_1-1}} \\ & + C_1(D_1)\cdot\log(v_3\ T^{1+\frac{1}{D_1+2}})\cdot(\log(Tv_3))^{-\frac{D_1+1}{D_1+2}} \\ & \cdot T^{1-\frac{1}{D_1+2}} \\ & + C_2\cdot T^{1-\frac{1}{D_1+2}}\cdot(\log(Tv_3))^{\frac{1}{D_1+2}} \end{aligned}$$

where $C_0(D_1) := 3QK(6Zv_1/v_2+2)L\frac{v_1 v_2^{-D_1}}{(\rho^{-1}-1)}$, $C_1(D_1) := 4QK(6Zv_1/v_2+2)\frac{v_2^{-D_1}}{Lv_1(\rho^{-1}-1)}$, $C_2 := K(6Zv_1/v_2+2)Lv_1$. We note that the condition $T > e/v_3$ is not restrictive since in a practical communication setting the number of packets sent (rounds) is large. For instance, we consider $T = 25,000$ in our experiments. In essence, we do not need to make any distributional assumptions on how contexts are generated, thus our performance bounds hold for any user mobility model. This is especially important in wireless communications since no model can perfectly capture the real-world mobility patterns of the users. Having said that, our performance analysis is not entirely agnostic to the sequence of contexts. The notion of approximate-optimality dimension captures both the structure of the reward function and the context sequence. This dimension can be significantly smaller than the actual context dimension in favourable cases. Nevertheless, it is never larger than the actual context dimension. This property showcases the power of adaptive discretization performed by our algorithm, which can accommodate any context arrival process by tuning exploration and exploitation based on the sequence of contexts observed thus far.

In the following corollary, we show that the growth of regret is the smallest when $|\mathcal{X}| < \infty$.

Corollary 1: Assume that the conditions in Theorem 1 hold and that $|\mathcal{X}| < \infty$. Fix $\epsilon > 0$ and $\delta \in (0,1)$. There exists $Q = Q(\mathcal{X},\mu,c) > 0$ (independent of T) such that the following bound holds for the regret of MUSIC with probability at least $1-\delta$:

$$\begin{aligned} R(T)\ \leq\ & C_0(\epsilon)\cdot T^{\frac{\epsilon-1}{\epsilon+2}}\cdot(\log(Tv_3))^{\frac{\epsilon+2}{\epsilon-1}} \\ & + C_1(\epsilon)\cdot\log(v_3\ T^{1+\frac{1}{\epsilon+2}})\cdot(\log(Tv_3))^{-\frac{\epsilon+1}{\epsilon+2}} \\ & \cdot T^{1-\frac{1}{\epsilon+2}} + C_2\cdot T^{1-\frac{1}{\epsilon+2}}\cdot(\log(Tv_3))^{\frac{1}{\epsilon+2}} \end{aligned}$$

where $C_0(\cdot)$, $C_1(\cdot)$, C_2 and c are defined in Theorem 1.

The results above characterize the worst-case regret of MUSIC.

Remark 4: Munkres' algorithm is an exact algorithm, which means that it will return an optimal solution to the approximate maximal matching problem formed by using parameter estimates (in our case upper confidence bounds) of the base arms. Since the computational complexity of Munkres' algorithm is polynomial in the input size (i.e., $O(n^3)$, where n is the dimension of the input matrix), using it allows our algorithm's performance to converge to that of the optimal dynamic assignment (i.e., achieve sublinear regret with respect to the optimal dynamic assignment) conditioned on the fact that our parameter estimates converge to the true mean base arm outcomes over time, without sacrificing computational efficiency. Without an exact algorithm, it will not be possible to obtain sublinear regret with respect to the optimal dynamic assignment. Thus, in the absence of an exact oracle, a relaxed notion of regret, called α-approximation regret [34] becomes a viable alternative that can be minimized.

B. Convergence and Complexity

Our regret bounds give an explicit bound on the rate of convergence. Theorem 1 provides a $\tilde{O}(T^{(\bar{D}+1)/(\bar{D}+2)+\epsilon})$ regret bound that holds for any $\epsilon > 0$, where $\bar{D}$ represents the approximate-optimality dimension which is less than or equal to the dimension of the context space. Note that

$$\frac{R(T)}{T} = \frac{\sum_{t=1}^{T} opt(\mu_t)}{T} - \frac{\sum_{t=1}^{T} u(\mu(\boldsymbol{x}_{t,S_t}))}{T}$$

represents the difference between expected total reward accumulated by the oracle and our algorithm. $R(T) \in \tilde{O}(T^{(\bar{D}+1)/(\bar{D}+2)+\epsilon})$ implies that $R(T)/T \in \tilde{O}(T^{(-1)/(\bar{D}+2)+\epsilon})$, which implies that $\lim_{T\to\infty} R(T)/T = 0$ for sufficiently small $\epsilon > 0$. Thus $\tilde{O}(T^{(-1)/(\bar{D}+2)+\epsilon})$ can be seen as the rate of convergence to the optimal average reward.

The time complexity of the algorithm is analyzed as follows: In every round t, the learner first identifies the active quantized contexts, and the complexity for this identification is $|\mathcal{A}_t|\cdot|\mathcal{L}_t|$. Then, after giving the indices as input to the Munkres' Oracle, it obtains S_t and identifies $\mathcal{P}_t$. Munkres' algorithm is known to have $O(n^3)$ time complexity, where n here is $E_t := \max\{N_t, \overline{C}\}$. Let $E = \max_{t\leq T} E_t$; thus, the complexity of running it is upper bounded by $O(E^3)$. Note that $|\mathcal{P}_t| \leq K_t$, and the algorithm iterates twice over the selected quantized contexts in $\mathcal{P}_t$ in order to update the indices and decide whether or not to refine the tree. Let $A = \max_{t\leq T}\max\{|\mathcal{A}_t|,|\mathcal{L}_t|\}$; thus, the algorithm's time complexity is $O\left(\sum_{t\leq T}\left(|\mathcal{A}_t|\cdot|\mathcal{L}_t| + E_t^3 + 2K_t\right)\right) = O(A^2\ T + E^3T + 2KT)$, which is a polynomial in the input parameters.

V. Experiments

Systematic simulations are carried out over an outdoor deployment with various system parameters to evaluate the performance of the proposed algorithm. We explicitly consider the impact of multi-user access and focus on available users surrounded by SBSs to maximize the overall system throughput.

A. Competitor Methods

- *Oracle*: Benchmark that knows the expected throughputs of all base associations, and has perfect knowledge of channel conditions. It chooses the optimal super association in every round.
- *CC-MAB*: The algorithm in [26] that takes into account both context and volatility of base associations. This algorithm partitions the context space uniformly at the beginning, and thus, is unable to adapt to non-uniform context arrivals, which is often the case in dynamic user-SBS association due to the fact that expected reward varies with distance and frequency.
- *Max Performance*: Traditional association algorithms that work for a given set of user-SBS positions determine

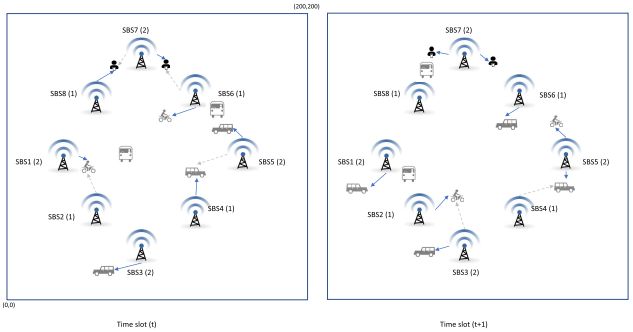

Fig. 4. There are 8 SBSs inside a squared outdoor region of $200m \times 200m$. Half of the SBSs have 2 channels and remaining half have 1 channel written in front of SBSs in brackets, and thus, SBSCs count is 12. User count and users positions are coming from different mobility models. Since, optimal association is changing at every time instance, an example of two consecutive time slots is shown.

associations by maximizing the received power or minimizing the path loss. Since the channel gains are unknown in our setting, to simulate a traditional association algorithm, we first estimate the throughput using a dry run for the generated context and channel states. Then, we calculate the best super association based on the estimated throughputs. We call this algorithm Max Performance. Max Performance returns a static super association, and thus, it ignores the contextual information.

- *CTS*: A combinatorial MAB algorithm based on Thompson Sampling (TS). CTS ignores the context, and thus, learns the best performing user-SBS associations on average.
- *CMAB*: A combinatorial MAB algorithm that uses upper confidence bounds. CMAB also ignores the context, and thus, learns the best performing user-SBS associations on average.
- *Random*: Selects user-SBSC associations randomly from the available base associations in each round. While this algorithm does not learn from its past experience, it serves as a benchmark to check if learning the optimal super association is a non-trivial task.

B. Simulation Setup

We consider a $200m \times 200m$ square outdoor area, and place an SBS network inside the area, which consists of $M = 8$ SBSs as shown in Fig. 4. Out of eight SBSs, four SBSs have 2 SBSCs each, whereas the remaining four SBS have 1 SBSC each, and thus, SBSCs count is 12. The transmission power and transmission rate for each SBS are kept to be fixed. We consider the following mobility models (MM) [57]:

1) Random walk (RW) on a grid: There are 7 users each of whom moves a fixed amount (0.5 meters) between two consecutive rounds. Movement is only possible in 4 directions: north, east, west, and south. At the end of each round, movement direction of each user is sampled uniformly at random from the set of possible directions.
2) Random trip (RT): There are 7 users. At the end of each round, velocity of each user is sampled uniformly at random from $[0, 1]$ (meters/round) and the angle (direction) of movement is sampled uniformly at random from $[0, 2\pi]$.
3) Group mobility (GM): There is one logical center of the group, which determines the movement of the group. At the end of each round, the center's velocity is sampled uniformly at random from $[0, 2]$ meters/round and its movement angle is sampled uniformly at random from $[0, 2\pi]$. There are 6 other group members. They move away at an angle sampled uniformly at random from $[0, 2\pi]$, and distance sampled uniformly at random from $[0, 0.5]$ meters from the logical center of the group.
4) Manhattan mobility in longitude only (MM-LG): The number of users is sampled uniformly at random from $\{0, \ldots, 6\}$, and selected users move in longitude only, whose positions are sampled from a uniformly from $[0, 200]$ meters, at fixed latitude positions.
5) Manhattan mobility in latitude only (MM-LT): The number of users is sampled uniformly at random from $\{0, \ldots, 6\}$, and selected users move in latitude only, whose positions are sampled from a uniformly from $[0, 200]$ meters, at fixed longitude positions.

TABLE II
SIMULATION PARAMETERS

Parameters	Value
M	8
K	12
Reference distance (d_0)	1 m
Transmit power	2.5 dBm
Noise level (σ^2)	-100 dBm
Modulation scheme	8-QAM
Carrier frequencies (f)	$\{28, 30, 35, 40, 42, 45, 48, 50, 73, 95, 100, 120\}$ GHz
$f_{\max}, d_{\max}$	150, 500
Pathloss exponent (η_{LOS}, η_{NLOS})	$2,\ 3 - (f/f^2_{\max})$

In total, there are 20 users always present in the system, who move according to Models 1 to 3. In each round, there can be up to 12 dynamic users, whose positions are determined by Models 4 to 5. Therefore, the maximum expected number of available SBSC-user associations is $12 \times 32 = 384$. This makes the maximum expected cardinality of super association set 8.64×10^{30}. The parameters used in the simulations are listed in Table II. For each experiment, we set the time horizon to $T = 25,000$ rounds and represent results averaged over 10 repetitions. For CC-MAB, we set $\alpha = 1$ and $h_T = \lceil T^{\frac{1}{4}} \rceil$. For MUSIC, we set $\delta = 0.01$, $v_1 = \sqrt{2}$, $v_2 = 1$, $\rho = 0.71$, and $Z = 2$. Since the channel statistics are unknown, we set $L = 2$. Note that the performance will improve when L is chosen with the knowledge of the channel statistics.

The random variable Y_{LOS} is modeled as a Rayleigh random variable with expected value as 10 dB for LOS and Y_{NLOS} with the expected value as 5 dB for NLOS in (7). The expected rewards are calculated over Rayleigh Fading channel similar to [50]. We transmit only one packet containing 1080 symbols in a time slot (round) for a base association, where modulation scheme dependent *bits-per-symbol* (bps) rate determines the number of transmission bits in that time slot. Similarly, the number of transmission bits in a time slot for a super association is equal to the summation

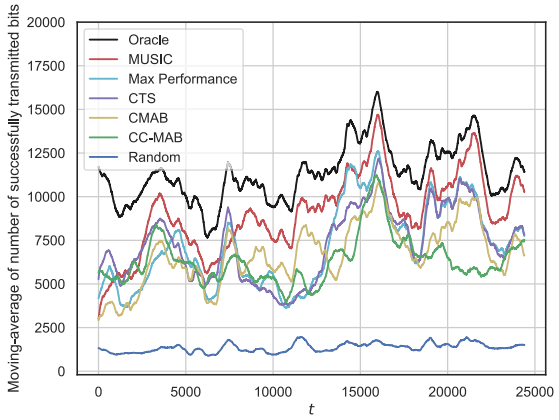

Fig. 5. Moving-average of number of successfully transmitted bits. Each value at time t is averaged over 10 repetitions of the experiment and 600 previous transmissions.

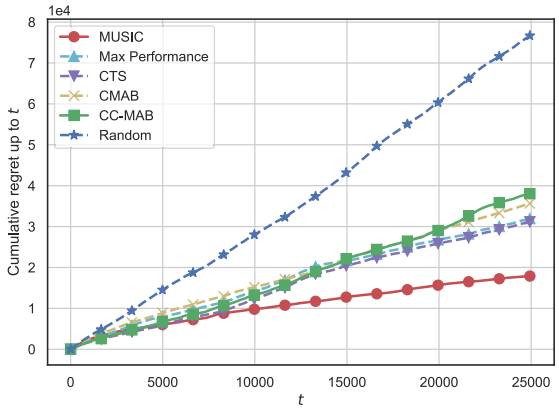

Fig. 6. Cumulative regrets of MUSIC, Max Performance, CTS, CMAB, CC-MAB and Random.

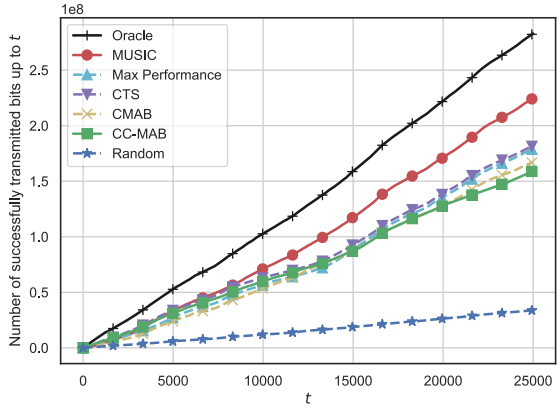

Fig. 7. Number of successfully transmitted bits up to t by Oracle, MUSIC, Max Performance, CTS, CMAB, CC-MAB and Random.

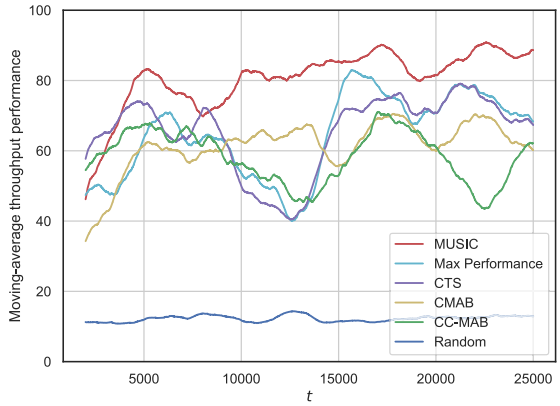

Fig. 8. Moving average throughput ratios of MUSIC, Max Performance, CTS, CMAB, CC-MAB and Random.

of transmission bits of base associations those constitute that super association.

C. Performance Comparison

Fig. 5 shows that the optimal successful transmission on average is around 11319 *bits*. MUSIC gets very close to this by achieving the successful transmission around 8992 *bits*. On the other hand, the successful transmissions achieved by Max Performance, CTS, CMAB, CC-MAB and Random are around 7168 *bits*, 7277 *bits*, 6686 *bits*, 6374 *bits*, 1354 *bits*, respectively. The achieved successful transmission in terms of bits by MUSIC is around 79% of the optimal strategy, and is around 25%, 24%, 34%, 41%, 564% higher than the successful transmission achieved by Max Performance, CTS, CMAB, CC-MAB and Random, respectively. In addition, we compare the throughput ratios of different algorithms in Fig. 8. We compute the moving average throughput ratio as the ratio of moving average of the throughput achieved by the corresponding algorithm and the throughput achieved by the oracle (calculated in percentage, where the theoretical maximum is 100%). It is evident from Fig. 8 that MUSIC is able to achieve more than 80% of the optimal throughput as time increases (note how the moving-average throughput ratio of MUSIC increases over time), while other algorithms lag far behind MUSIC.

Fig. 6 compares the cumulative regrets of Max Performance, CTS, CMAB, CC-MAB and Random. From this figure, it is evident that MUSIC converges to the optimal association faster than its competitors by accumulating a significantly smaller regret. Similarly, Fig. 7 compares the number of successfully transmitted bits up to round t of Oracle, MUSIC, Max Performance, CTS, CMAB, CC-MAB and Random. The number of successfully transmitted bits of MUSIC are closest to Oracle, while other algorithms have much lower successful transmissions. It is observed that non-contextual algorithms and CC-MAB cannot learn fast enough in this setting. CC-MAB learns slowly because it uses uniform partitioning and contexts arrive non-uniformly. With uniform partitioning, a considerably larger time horizon is required to learn enough. On the other hand, MUSIC is able to perform much better by adaptively zooming into the regions with dense context arrivals.

D. Position Dependent Unknown Blockage and Role of the Context

As discussed in Remark 1, we choose a position dependent blockage model and position dependent path loss, where not only the blockage probabilities are unknown but also vary with position of a given user-SBS association. Therefore, we add angle between the user and the SBS as an additional context

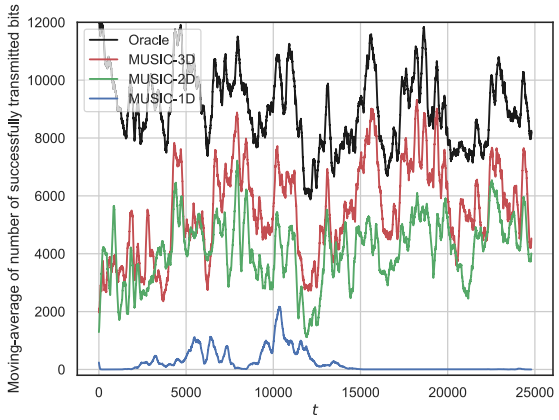

Fig. 9. Moving-average of number of successfully transmitted bits, each value at time t is averaged over 10 repetitions of the experiment and 600 previous transmissions.

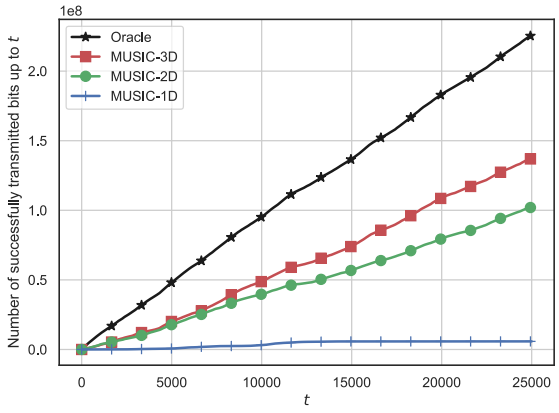

Fig. 10. Number of successfully transmitted bits up to t by MUSIC-1D, MUSIC-2D and MUSIC-3D.

dimension which is normalized in $[0, 1]$. The expected reward is updated as the product of the earlier frequency-distance tuple dependent expected reward and the 4th power of normalized angle (i.e., $\tilde{\mu} = (angle)^4 \times \mu_{LOS}$). The received average SNR is also considered as a function of angle (i.e., $SNR{=}10{+}5{\times}\, angle$ dBs). Therefore, not only the path loss but also the reward function, which depends on transmission frequency and distance, depends on position in the considered scenario. We also study the effect of context choice on the performance of the MUSIC. Since the expected reward is dependent on the frequency, distance, and angle, we compare three variants of MUSIC: i) MUSIC-1D, which considers only the frequency as 1-dimensional (1-D) context, ii) MUSIC-2D which considers the frequency and distance as 2-D context, and iii) MUSIC-3D, which considers the frequency, distance and angle as 3-D context. The parameters are context dependent and are set as $L = D$, and $\rho = 2^{-1/D}$, where D is the dimension of context. It is shown in Fig. 9 and Fig. 10 that when system throughput is dependent on frequency, distance and angle, exploiting all the contextual information helps to improve the performance.

VI. Conclusion

We considered the association problem between dynamic users and small base stations over unknown wireless channels. We proposed an algorithm, called MUSIC, that performs adaptive discretization of the context space and achieves $\tilde{O}(T^{(\tilde{D}+1)/(\tilde{D}+2)+\epsilon})$ regret for any $\epsilon > 0$, where $\tilde{D}$ represents the approximate-optimality dimension related to the context space. We performed experiments on mmWave channel models with different forms of user mobility, and showed that MUSIC significantly improves the performance without requiring CSI or knowledge of channel statistics.

Appendix

A. Approximate-Optimality Dimension

Definition 3:

- *A subset $\mathcal{X}_2$ of $\mathcal{X}$ is called r-separated if for any $x_1, x_2 \in \mathcal{X}_2$ such that $x_1 \neq x_2$, we have $\bar{d}(x_1, x_2) \geq r$. The cardinality of the largest such set is called the r-packing number of $\mathcal{X}$ with respect to $\bar{d}$, and is denoted by $M(\mathcal{X}, \bar{d}, r)$. Equivalently, the r-packing number of $\mathcal{X}$ is the maximum number of disjoint $\bar{d}$-balls of radius r that are contained in $\mathcal{X}$.*
- *For any $t \geq 1$, let $\overline{\mathcal{X}}_t = \mathcal{X}^{K_t}$. Given $r > 0$ and $f : \mathbb{R}^+ \to \mathbb{R}^+$, we define*

$$\mathcal{X}_{f(r)} := \{x \in \mathcal{X} : opt(\mu_t) - u(\mu(\mathbf{x})) \leq f(r), \text{ for some } \mathbf{x} \in \overline{\mathcal{X}}_t \text{ and } t \geq 1 \text{ such that } x \in \mathbf{x}\}$$

to be an $(f(r), \mu)$-optimal set. Let $M(\mathcal{X}_{f(r)}, \bar{d}, r)$ be its r-packing number. We define the (f, μ)-optimality dimension $D^\mu(f)$ associated with $\mathcal{X}_{f(r)}$ and μ as follows:

$$D^\mu(f) = \max\left\{0, \limsup_{r \to 0} \frac{\log(M(\mathcal{X}_{f(r)}, \bar{d}, r))}{\log(r^{-1})}\right\}.$$

Since $\mathcal{X}_{f(r)} \subseteq \mathcal{X}$, we have $D^\mu(f) \leq 2$. Based on the arrival patterns of the users, and variations in the contexts, $D^\mu(f)$ can be as small as 0. Our worst-case regret bounds will depend on $D^\mu(f)$.

B. Proof of Lemma 1

Let $h(z) = e^{-z}$. Note that $h'(z) = -e^{-z}$, thus we have $|h'(z)| \leq 1$ for $z \in (0, \infty)$. This implies that $|h(z) - h(z')| \leq |z - z'|$ for $z \in (0, \infty)$. Similarly, for $l(z) = 1 - e^{-z}$, we have $|l(z) - l(z')| \leq |z - z'|$ for $z \in (0, \infty)$.

We have $x = [\frac{f}{f_{\max}}, \frac{d}{d_{\max}}]$, $w_{LOS}(x) = \frac{1}{f^2\ d^{\eta_{LOS}}}$, and we assume $\eta_{LOS}, f, d \geq 1$. Let $\eta = \eta_{LOS}$ and $q(x) = \frac{yG_{LOS}}{f^2\ d^\eta}$ for $y \geq 0$. Observe that $|\frac{\partial q(x)}{\partial f}| = \frac{2\ yG_{LOS}}{f^3\ d^\eta} \leq \frac{2\ yG_{LOS}}{d^\eta}$ for $f \geq 1$, and thus, $|\frac{yG_{LOS}}{f^2\ d^\eta} - \frac{yG_{LOS}}{f'^2\ d^\eta}| \leq \frac{2\ yG_{LOS}}{d^\eta}|f - f'| \leq 2\ yG_{LOS}\ |f - f'|$. Similarly, $|\frac{\partial q(x)}{\partial d}| = \frac{\eta y G_{LOS}}{f^2\ d^{\eta+1}} \leq \eta y G_{LOS}$ for $f, d, \eta \geq 1$, and thus, $|\frac{yG_{LOS}}{f'^2\ d^\eta} - \frac{yG_{LOS}}{f'^2 (d')^\eta}| \leq \eta_{LOS} G_{LOS} y\ |d - d'|$.

We have

$$\begin{aligned}
&|h \circ q(x) - h \circ q(x')| \\
&= |h(q(x)) - h(q(x'))| \\
&\leq |q(x) - q(x')| \\
&\leq |\frac{yG_{LOS}}{f^2\ d^{\eta_{LOS}}} - \frac{yG_{LOS}}{f'^2\ d^{\eta_{LOS}}}| + |\frac{yG_{LOS}}{f'^2\ d^{\eta_{LOS}}} - \frac{yG_{LOS}}{f'^2 (d')^{\eta_{LOS}}}| \\
&\leq 2\ yG_{LOS}|f - f'| + y\eta_{LOS} G_{LOS}|d - d'|.
\end{aligned}$$

We write

$$\begin{aligned}
&|\mu_{LOS}(x) - \mu_{LOS}(x')| \\
&\quad = r\Bigg| \int_y \Big(\exp\left(-G_{LOS} w_{LOS}(x') y\right) \\
&\qquad - \exp\left(-G_{LOS} w_{LOS}(x) y\right)\Big) f_{Y_{LOS}}(y) dy \Bigg| \\
&\quad \leq \Big(2\, G_{LOS} r |f - f'| + \eta_{LOS} G_{LOS} r |d - d'|\Big) \\
&\qquad \times \int_y y f_{Y_{LOS}}(y) dy \\
&\quad \leq L^f_{LOS} |f - f'| + L^d_{LOS} |d - d'|.
\end{aligned}$$

where $L^f_{LOS} = 2\ G_{LOS} r E[Y_{LOS}]$, and $L^d_{LOS} = \eta_{LOS} G_{LOS} r E[Y_{LOS}]$. Similarly, $|\mu_{NLOS}(x) - \mu_{NLOS}(x')| \leq L^f_{NLOS}|f - f'| + L^d_{NLOS}|d - d'|$, where $L^f_{NLOS} = 2\ G_{NLOS} r E[Y_{NLOS}]$, and $L^d_{NLOS} = \eta_{NLOS} G_{NLOS} r E[Y_{NLOS}]$.

Furthermore, we use Theorem 12.1 and 12.4 in [58, Chapter 12], which provide the fact that the function generated by the linear combination of multiple Lipschitz functions or the product of bounded Lipschitz functions is also Lipschitz. Using (4), we have $P_{LOS}(d) = \min(1, \zeta_0\ e^{-\delta_0\ d}) \zeta_{LOS} e^{-\delta_{LOS} d}$, and by using the fact that $\min(h_1, h_2) = \frac{h_1 + h_2 - |h_1 - h_2|}{2}$, we have $|P_{LOS}(d) - P_{LOS}(d')| \leq (\zeta_0 \delta_0 + \zeta_{LOS} \delta_{LOS}) |d - d'|$. Similarly, $|P_{NLOS}(d) - P_{NLOS}(d')| \leq (\zeta_0 \delta_0 + \zeta_{NLOS} \delta_{NLOS}) |d - d'|$. Therefore, by using $L = 2 \cdot \max\{f_{\max}\big(L^f_{LOS} + L^f_{NLOS}\big), d_{\max}\big(L^d_{LOS} + L^d_{NLOS} + 2\zeta_0\delta_0 + \zeta_{LOS}\delta_{LOS} + \zeta_{NLOS}\delta_{NLOS}\big)\big)\}$, and $\bar{d}(x, x') = |\frac{d}{d_{\max}} - \frac{d'}{d_{\max}}| + |\frac{f}{f_{\max}} - \frac{f'}{f_{\max}}|$, we have, $|\mu(x) - \mu(x')| \leq L \bar{d}(x, x')$.

C. Technical Lemmas Used in the Proofs

Lemmas given in this section are analogous to the lemmas in [34]. Thus, their proofs are omitted for brevity. $\mathcal{Z}_t$ includes parent quantized contexts as well while $\mathcal{L}_t$ does not include parent quantized contexts. Let $\mathcal{Z}_t$ denote the set of quantized contexts created by round t. The following lemma defines a good event, which holds with a high probability. We will show that the regret will be small when this good event holds. Henceforth, we will refer to quantized contexts as nodes for brevity.

Lemma 2: Given $\delta \in (0, 1)$, let

$$\begin{aligned}
&c_t(x^{(h,i)}) \\
&= \sqrt{\frac{(1 + B_t(x^{(h,i)}))}{(B_t(x^{(h,i)}))^2} \left(1 + 2\log\left(\frac{KT(1 + B_t(x^{(h,i)}))^{1/2}}{\delta}\right)\right)},
\end{aligned}$$

defined as in (9)*, for any node $x^{(h,i)}$ and time $t \geq 1$. Then, given the event $\mathcal{F} = \big\{\forall t \leq T, \forall x^{(h,i)} \in \mathcal{Z}_t : |\hat{\mu}_t(x^{(h,i)}) - \mu(x^{(h,i)})| \leq c_t(x^{(h,i)}) + L v_1 \rho^h\big\}$, we have $\mathbb{P}\{\mathcal{F}\} \geq 1 - \delta$.*

The next result gives high probability bounds of the difference between the index and the mean of a given node.

Lemma 3: Consider that event $\mathcal{F}$ happens. Then, if at round t, the node $x^{(h,i)} \in \mathcal{P}_t$ is not expanded by the algorithm, we have $\overline{\mu}_t(x^{(h,i)}) - \mu(x^{(h,i)}) \leq (5 Z v_1/v_2 + 1) L v_1 \rho^h$. Moreover, a node $x^{(h,i)}$ may be selected by the algorithm no more than q_h times before it is expanded, where $q_h < 3 + \frac{4}{(L v_1 \rho^h)^2} \log\left(\frac{T v_3}{\rho^h}\right)$ with $v_3 = \frac{K\sqrt{e}}{L v_1 \delta}$.

Next, we give a high probability upper bound on the true mean of any given arm.

Lemma 4: Consider that event $\mathcal{F}$ happens. Then, we have $\forall t \leq T$ and $a \in \mathcal{A}_t$, $\overline{\mu}_t(x_{t,a}) \geq \mu(x_{t,a})$.

Define the suboptimality gap in round t as $\Delta(S_t) := opt(\mu_t) - u(\mu(\mathbf{x}_{t,S_t}))$. The next lemma gives a bound on $\Delta(S_t)$.

Lemma 5: Under the established assumptions, if event $\mathcal{F}$ holds, then in any round t, we have $\Delta(S_t) \leq K(6 Z v_1/v_2 + 2) L v_1 \rho^{H_t}$, where $H_t = \min\{H_{t,k} : 1 \leq k \leq K_t\}$.

Lastly, we will make use of the following fact from [34] when upper bounding the cardinality of the set of nodes from which the algorithm selects.

Lemma 6: Fix $\kappa > 0$. Let $\bar{D} = D^{\mu}(f)$ and $f(r) = cr$ for a given $c > 0$. Fix $D_1 > \bar{D}$. Then, there exists a constant Q, such that for all $r \leq v_2$ we have $M(\mathcal{X}_{cr}, \overline{d}, r) \leq Q r^{-D_1}$.

D. Proof of Theorem 1

Assume event $\mathcal{F}$ happens. Thus, the condition of Lemma 2 is satisfied. Lemma 5 indicates that the context of the kth selected base association in round t is in a $(K(6 Z v_1/v_2 + 2) L v_1 \rho^{H_t}, \mu)$-optimal set, i.e., $\tilde{x}_{t,k} \in \mathcal{X}_{K(6 Z v_1/v_2 + 2) L v_1 \rho^{H_t}}$. Thus, the regret is bounded as

$$R(T) \leq \sum_{t \leq T} K(6 Z v_1/v_2 + 2) L v_1 \rho^{H_t}\ .$$

Fix a positive number H (exact value will be specified later). Let

$$R_1(T) := \sum_{t \leq T; H_t < H} K(6 Z v_1/v_2 + 2) L v_1 \rho^{H_t}$$

represent the regret coming from base associations with node levels smaller than H and

$$R_2(T) := \sum_{t \leq T; H_t \geq H} K(6 Z v_1/v_2 + 2) L v_1 \rho^{H_t}$$

represent the regret coming from base associations with node levels larger than or equal to H. We have $R(T) \leq R_1(T) + R_2(T)$. Below, we bound $R_1(T)$:

$$\begin{aligned}
&R_1(T) \\
&\leq \sum_{h=0}^{H-1} |\mathcal{X}^h \cap \mathcal{X}_{K(6 Z v_1/v_2 + 2) L v_1 \rho^h}| \\
&\quad \cdot K(6 Z v_1/v_2 + 2) L v_1 \rho^h \cdot q_h \qquad (12) \\
&\leq \sum_{h=0}^{H-1} M(\mathcal{X}_{K(6 Z v_1/v_2 + 2) L v_1 \rho^h}, \overline{d}, v_2 \rho^h) \\
&\quad \cdot K(6 Z v_1/v_2 + 2) L v_1 \rho^h \cdot q_h \qquad (13) \\
&\leq \sum_{h=0}^{H-1} Q \cdot (v_2 \rho^h)^{-D_1} K(6 Z v_1/v_2 + 2) L v_1 \rho^h \\
&\quad \cdot \left(3 + \frac{4}{(L v_1 \rho^h)^2} \log\left(\frac{T v_3}{\rho^h}\right)\right) \qquad (14)
\end{aligned}$$

$$
\begin{aligned}
&= 3QK(6Zv_1/v_2+2)Lv_2^{-D_1}v_1\sum_{h=0}^{H-1}\rho^{-h(D_1-1)} \\
&\quad + 4QK(6Zv_1/v_2+2)\frac{v_2^{-D_1}}{Lv_1}\sum_{h=0}^{H-1}\rho^{-h(D_1+1)}\log\left(\frac{Tv_3}{\rho^h}\right) \\
&\leq 3QK(6Zv_1/v_2+2)L\frac{v_1v_2^{-D_1}}{(\rho^{-1}-1)}\rho^{-H(D_1-1)} \\
&\quad + 4QK(6Zv_1/v_2+2)\frac{v_2^{-D_1}}{Lv_1(\rho^{-1}-1)} \\
&\quad \times \log\left(\frac{Tv_3}{\rho^H}\right)\rho^{-H(D_1+1)} \ . \qquad (15)
\end{aligned}
$$

We see that (12) holds since for a certain level h, in T rounds, the learner may have selected up to $|\mathcal{X}^h \cap \mathcal{X}_{K(6Zv_1/v_2+2)Lv_1\rho^h}|$ nodes, and any of them up to q_h times. Any of these nodes contributes to the regret with a maximum amount of $K(6Zv_1/v_2+2)Lv_1\rho^h$. (13) follows from the definition of $v_2\rho^h$-packing number and the fact that any two nodes in $\mathcal{X}^h$ are at least $v_2\rho^h$ apart. In (14), we use Lemma 6, since $v_2\rho^h \leq v_2$, for any $h \geq 0$, and the fact that $q_h < \left(3+\frac{4}{(Lv_1\rho^h)^2}\log\left(\frac{Tv_3}{\rho^h}\right)\right)$ in MUSIC.

$R_2(T)$ can be bounded as:

$$
R_2(T) \leq TK(6Zv_1/v_2+2)Lv_1\rho^H \ . \qquad (16)
$$

Summing the bounds in and (15) and (16), we get

$$
\begin{aligned}
R(T) \ &\leq C_0(D_1)\rho^{-H(D_1-1)} \\
&\quad + C_1(D_1)\rho^{-H(D_1+1)}\log\left(\frac{Tv_3}{\rho^H}\right) + C_2\rho^H T \, . \qquad (17)
\end{aligned}
$$

Now, let $H = -\log_\rho\left(\frac{T}{\log(Tv_3)}\right)^{\frac{1}{D_1+2}}$, which gives $\rho^{-H} = \left(\frac{T}{\log(Tv_3)}\right)^{\frac{1}{D_1+2}}$, $\rho^H = \left(\frac{\log(Tv_3)}{T}\right)^{\frac{1}{D_1+2}}$ and

$$
\begin{aligned}
\log\left(\frac{Tv_3}{\rho^H}\right) &= \log\left((Tv_3)\left(\frac{T}{\log(Tv_3)}\right)^{\frac{1}{D_1+2}}\right) \\
&\leq \log\left(v_3\ T^{1+\frac{1}{D_1+2}}\right) \ .
\end{aligned}
$$

For the last inequality we use the fact that $T > e/v_3$. Finally, we obtain the main result by substituting the above bounds in (17).

E. Proof of Corollary 1

As $|\mathcal{X}| < \infty$, the r-packing number of $\mathcal{X}_{cr}$ for $c = K(6Zv_1/v_2+2)(v_1/v_2)L$ cannot exceed $|\mathcal{X}|$. Thus, $\bar{D} = \limsup_{r\to 0}\frac{\log(M(\mathcal{X}_{cr},\overline{d},r))}{\log(r^{-1})} \leq \limsup_{r\to 0}\frac{|\mathcal{X}|}{\log(r^{-1})} = 0$. Then, by Lemma 6 there exists some positive constant Q such that, we have $M(\mathcal{X}_{cr},\overline{d},r) \leq Qr^{-\epsilon}$. for all $\epsilon > 0$. The rest of the proof follows the same arguments as the proof of Theorem 1. We first fix an H, and then, separate the regret due to nodes with levels less than H (term $R_1(T)$) and greater than or equal to H (term $R_2(T)$). Then, we bound $R_1(T)$ and $R_2(T)$ as in Theorem 1 and sum these bounds by setting $H = -\log_\rho\left(\frac{T}{\log(Tv_3)}\right)^{\frac{1}{\epsilon+2}}$.

REFERENCES

[1] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[2] K. Akkarajitsakul, E. Hossain, D. Niyato, and D. I. Kim, "Game theoretic approaches for multiple access in wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 3, pp. 372–395, 3rd Quart., 2011.

[3] H. W. Kuhn, "The hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[4] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.

[5] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[6] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2014, pp. 1096–1104.

[7] Y. Zhou, C. Shen, and M. van der Schaar, "A non-stationary online learning approach to mobility management," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1434–1446, Feb. 2019.

[8] A. Ganti, T. E. Klein, and M. Haner, "Base station assignment and power control algorithms for data users in a wireless multiaccess framework," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2493–2503, Sep. 2006.

[9] E. Yaacoub and Z. Dawy, "A survey on uplink resource allocation in OFDMA wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 322–337, 2nd Quart., 2012.

[10] S. Sadr, A. Anpalagan, and K. Raahemifar, "Radio resource allocation algorithms for the downlink of multiuser OFDM communication systems," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 3, pp. 92–106, 3rd Quart., 2009.

[11] J. Jang and K. Bok Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171–178, Feb. 2003.

[12] W. Yu, G. Ginis, and J. M. Cioffi, "Distributed multiuser power control for digital subscriber lines," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 5, pp. 1105–1115, Jun. 2002.

[13] I. Bistritz and A. Leshem, "Asymptotically optimal resource block allocation with limited feedback," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 34–46, Jan. 2019.

[14] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[15] S. A. Ramprashad and G. Caire, "Cellular vs. network MIMO: A comparison including the channel state information overhead," in *Proc. IEEE 20th Int. Symp. Pers., Indoor Mobile Radio Commun.*, Sep. 2009, pp. 878–884.

[16] M. Feng, S. Mao, and T. Jiang, "BOOST: Base station ON-OFF switching strategy for energy efficient massive MIMO HetNets," in *Proc. IEEE 35th Annu. Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.

[17] L. Liu, S. Zhang, and R. Zhang, "CoMP in the sky: UAV placement and movement optimization for multi-user communications," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5645–5658, Aug. 2019.

[18] G. Athanasiou, T. Korakis, O. Ercetin, and L. Tassiulas, "A cross-layer framework for association control in wireless mesh networks," *IEEE Trans. Mobile Comput.*, vol. 8, no. 1, pp. 65–80, Jan. 2009.

[19] M. A. Qureshi and C. Tekin, "Fast learning for dynamic resource allocation in AI-enabled radio networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 1, pp. 95–110, Mar. 2020.

[20] R. Combes and A. Proutiere, "Dynamic rate and channel selection in cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 5, pp. 910–921, May 2015.

[21] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

[22] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3–4, pp. 285–294, Dec. 1933.

[23] A. Slivkins, "Contextual bandits with similarity information," *J. Mach. Learn. Res.*, vol. 15, pp. 2533–2568, Jan. 2014.

[24] C. Tekin and M. van der Schaar, "Active learning in context-driven stream mining with an application to image mining," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3666–3679, Nov. 2015.

[25] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Proc. 20th Neural Inf. Process. Syst.*, 2007, pp. 817–824.
[26] L. Chen, J. Xu, and Z. Lu, "Contextual combinatorial multi-armed bandits with volatile arms and submodular reward," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 3247–3256.
[27] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 151–159.
[28] A. Huyuk and C. Tekin, "Analysis of Thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms," in *Proc. 22nd Int. Conf. Artif. Intell. Stat.*, 2019, pp. 1322–1330.
[29] Z. Bnaya, R. Puzis, R. Stern, and A. Felner, "Volatile multi-armed bandits for guaranteed targeted social crawling," in *Proc. 27th AAAI Conf. Artif. Intell. Workshops*, 2013, pp. 15–21.
[30] D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal, "Mortal multi-armed bandits," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 273–280.
[31] I. Bistritz and A. Leshem, "Distributed multi-player bandits-a game of thrones approach," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 7222–7232.
[32] H. Tibrewal, S. Patchala, M. K. Hanawal, and S. J. Darak, "Distributed learning and optimal assignment in multiplayer heterogeneous networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2019, pp. 1693–1701.
[33] A. Magesh and V. V. Veeravalli, "Multi-player multi-armed bandits with non-zero rewards on collisions for uncoordinated spectrum access," 2019, *arXiv:1910.09089*. [Online]. Available: http://arxiv.org/abs/1910.09089
[34] A. Nika, S. Elahi, and C. Tekin, "Contextual combinatorial volatile multi-armed bandit with adaptive discretization," in *Proc. 23rd Int. Conf. Artif. Intell. Stat.*, 2020, pp. 1486–1496.
[35] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. Neural Inf. Process. Syst.*, 2011, pp. 2312–2320.
[36] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "X-armed bandits," *J. Mach. Learn. Res.*, vol. 12, pp. 1655–1695, Jan. 2011.
[37] H. U. Sokun, E. Bedeer, R. H. Gohary, and H. Yanikomeroglu, "Optimization of discrete power and resource block allocation for achieving maximum energy efficiency in OFDMA networks," *IEEE Access*, vol. 5, pp. 8648–8658, 2017.
[38] J. Chen, R. Berry, and M. Honig, "Limited feedback schemes for downlink OFDMA based on sub-channel groups," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1451–1461, Oct. 2008.
[39] S. Sanayei and A. Nosratinia, "Opportunistic downlink transmission with limited feedback," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 4363–4372, Nov. 2007.
[40] H. Lou *et al.*, "Method and apparatus for supporting coordinated orthogonal block-based resource allocation (COBRA) operations," U.S. Patent 13 826 402, Oct. 31, 2013.
[41] J. Sharony and A. C. Sevdinoglou, "On-line distributed TDMA/FDMA/CDMA link assignment in mobile radio networks with flexible directivity," U.S. Patent 5 742 593, Apr. 21, 1998.
[42] T. S. Rappaport, *Wireless Communications: Principles and Practice*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.
[43] J. S. Seybold, *Introduction to RF Propagation*. Hoboken, NJ, USA: Wiley, 2005.
[44] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2nd Quart., 2018.
[45] S. Sun *et al.*, "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5G wireless communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 2843–2860, May 2016.
[46] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—With a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.
[47] M. R. Akdeniz *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.
[48] S.-C. Lin and I. F. Akyildiz, "Dynamic base station formation for solving NLOS problem in 5G millimeter-wave communication," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.
[49] M. Di Renzo, "Stochastic geometry modeling and analysis of multi-tier millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5038–5057, Sep. 2015.
[50] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746–1755, Sep. 2004.
[51] O. Naparstek, S. M. Zafaruddin, A. Leshem, and E. A. Jorswieck, "Distributed energy efficient channel allocation," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 1152–1166, Dec. 2019.
[52] S. Shekhar and T. Javidi, "Gaussian process bandits with adaptive discretization," *Electron. J. Statist.*, vol. 12, no. 2, pp. 3829–3874, 2018.
[53] I. Bistritz, T. Baharav, A. Leshem, and N. Bambos, "My fair bandit: Distributed learning of max-min fairness with multi-player bandits," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 930–940.
[54] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1617–1626.
[55] M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 325–333.
[56] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.
[57] P. Santi, *Mobility Models for Next Generation Wireless Networks: Ad Hoc, Vehicular and Mesh Networks*. Hoboken, NJ, USA: Wiley, 2012.
[58] K. Eriksson, D. Estep, and C. Johnson, *Applied Mathematics: Body and Soul: Derivatives and Geometry in IR3*, vol. 1. Cham, Switzerland: Springer, 2013.

Muhammad Anjum Qureshi (Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from UET, Taxila, Pakistan, in 2005, the master's degree from CASE, Islamabad, Pakistan, in 2010, and the Ph.D. degree with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey, in 2020, under the supervision of Dr. C. Tekin. His research interests include machine learning, wireless communications, and multiarmed bandit problems. He received the Alper Atalay Award for Best Article in IEEE-SIU 2017 and the Third Best Student Paper Award in IEEE-SIU 2018.

Andi Nika received the B.Sc. degree in mathematics from the University of Tirana, Albania, in 2015, and the master's degree in mathematics from Bilkent University, Ankara, Turkey, in 2018, where he is currently pursuing the second master's degree with the Department of Electrical and Electronics Engineering, under the supervision of Dr. C. Tekin. His research interests include machine learning, active learning, multiarmed bandits, and Bayesian optimization.

Cem Tekin (Senior Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2008, and the M.S.E. degree in electrical engineering: systems, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering: systems from the University of Michigan, Ann Arbor, MI, USA, in 2010, 2011, and 2013, respectively. From February 2013 to January 2015, he was a Post-Doctoral Scholar with the University of California, Los Angeles, CA, USA. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey. His research interests include cognitive communications, reinforcement learning, multiarmed bandit problems, and multiagent systems. He received numerous awards, including the Fred W. Ellersick Award for the best article in MILCOM 2009 and the Distinguished Young Scientist (BAGEP) Award of the Science Academy Association of Turkey in 2019.