Joint AP Probing and Scheduling: A Contextual Bandit Approach

Tianyi Xu[a], Ding Zhang[b], Parth H. Pathak[b], Zizhan Zheng[a]
[a] Department of Computer Science, Tulane University, New Orleans, LA, USA
[b] Department of Computer Science, George Mason University, Fairfax, VA, USA

***Abstract*—We consider a set of APs with unknown data rates that cooperatively serve a mobile client. The data rate of each link is i.i.d. sampled from a distribution that is unknown a priori. In contrast to traditional link scheduling problems under uncertainty, we assume that in each time step, the device can probe a subset of links before deciding which one to use. We model this problem as a contextual bandit problem with probing (CBwP) and present an efficient algorithm. We further establish the regret of our algorithm for links with Bernoulli data rates. Our CBwP model is a novel extension of the classic contextual bandit model and can potentially be applied to a large class of sequential decision-making problems that involve joint probing and play under uncertainty.**

***Index Terms*—joint probing and play, multi-armed bandits**

I. Introduction

Developing efficient resource allocation algorithms plays a central role in wireless networks research. Over the years, elegant solutions with performance guarantees have been designed for various tasks, including link scheduling, rate adaptation, power allocation, routing, and network utility optimization. A key assumption in many of these approaches is that the decision maker has perfect prior knowledge about the channel conditions. However, obtaining accurate channel conditions often requires substantial measurements, which can be very time consuming for both outdoor environments and dense indoor deployments. This is especially the case for mobile networks where the capacity of a link varies significantly over the locations of devices and environmental factors such as interference from other links and blockages. Thus, traditional approaches based on fixed channel conditions cannot obtain expected performance in unknown/uncertain environments or quickly adapt to changing environments.

To cope with the various uncertainty in network resource allocation and obtain adaptive scheduling policies, learning based approaches have been intensively studied recently. In particular, various online learning based algorithms have been developed for link scheduling [1], [2], rate adaptation [3] and beam selection [4], just to name a few. These works consider the challenging setting where the capacity of a link follows an unknown distribution that can only be sampled when the link is activated (i.e., the bandit feedback). Instead of using an offline learning approach with separated data collection and decision making stages, they adopt a multi-armed bandit based online learning framework that integrates exploration and exploitation. By carefully balancing the two aspects, they obtain no-regret adaptive policies with long-run performance approaching what can be achieved by the best offline policies that have prior knowledge on channel conditions.

In many real settings, a decision maker may obtain additional observations beyond the pure bandit feedback. For example, in the next generation millimeter-wave 802.11ad/ay WLANs, beamforming can be used to infer the real-time link quality before a scheduling decision is made. However, beamforming between all APs and clients can be time consuming for a densely deployed WLAN. Thus it is more realistic to assume that only a subset of APs can be selected for beamforming in each round, which reduces channel uncertainty but does not completely mitigate it. In this case, it is crucial to jointly optimize AP selection for beamforming (probing) and link scheduling for serving clients (play).

In this paper, we present a novel extension of the bandit learning framework to incorporate joint probing and play. We assume that before the decision maker chooses an arm to play in each round, it can probe a subset of arms and observe their rewards (in that round). The decision maker then picks an arm to play according to the observations obtained in the probing stage and historical data. Our framework can be directly applied to the joint beamforming and scheduling problem when multiple APs collaboratively serve a single client (detailed system model in Sec. III). Given that the data rate a client can obtain from an AP is highly correlated with the client location, we consider a contextual bandit model and treat the client location (or an approximation of it) as the context and learn a context-dependent joint probing and play policy.

To solve the problem, we first derive useful structural properties of the offline optimal solution and then develop an online learning algorithm by extending the contextual zooming algorithm in [4]. We further establish the regret bound of our algorithm in the special case when the reward distributions are Bernoulli. We apply our framework to the joint beamforming and scheduling problem in 802.11ad WLANs where a set of APs collaboratively serve a single mobile client. Simulations using real data traces demonstrate the efficiency of our solution.

Our bandit learning model and its extensions can potentially be applied to a large body of sequential decision making problems that involve joint probing and play under uncertainty. For example, by integrating probing with combinatorial multi-armed bandits where the decision maker can pick multiple arms to play, we can model the joint beamforming and scheduling problem in the more general multi-AP multi-client setting. As another example, consider the problem of finding the shortest path between a source and a destination in a road network with unknown traffic, where a path searcher

can query a travel server to obtain hints of real time travel latency [5]. Since each query consumes server resources and incurs delay, the path searcher can only make a limited number of queries before picking a path. Further, the path searcher may utilize contextual information such as the current time to assist decision making. This problem can again be modeled as a contextual combinatorial bandit problem with probing.

II. Related work

The classic multi-armed bandit (MAB) model provides a clean framework for studying the exploration vs. exploitation tradeoff in sequential decision making under uncertainty. Since the seminal work of Lai and Robbins [6], MAB and its variants have been intensively studied [7]–[9] and applied to various domains including wireless resource allocation. In particular, a combinatorial sleeping multi-armed bandit model with fairness constraints is considered in [2], which has been used to model single AP scheduling where multiple clients compete for sending packets to the common AP. In [3], the problem of link rate selection for a single wireless link is considered and a constrained Thompson sampling algorithm is developed to exploit the structural property that a higher data rate is associated with a lower transmission success probability. In [1], online learning based scheduling for general ad hoc wireless networks with unknown channel statistics is considered. The classic greedy maximal matching based algorithm is extended by using UCB-based link weights. The work that is closest to ours is [10], where a contextual multi-armed bandit algorithm is applied to the beam selection problem in mmWave vehicular systems. However, none of the above work considers the joint probing and scheduling problem as we consider in this paper.

Probing strategies for independent distributions have been studied in various domains including database query optimization [11]–[13] and wireless communication [14]. A common setting is that given a set of random variables with *known* distributions, a limited number of probes (observations) can be made about these distributions. A selection decision is then made according to the observations. This corresponds to the offline problem in our setting. Various objective functions have been considered including maximizing the largest value found minus the total probing cost spent. Since the general problem is NP-hard, various approximation algorithms have been developed [15]. More recently, adaptive probing strategies have been studied for shortest path routing [5] and when the probing cost varies across different alternatives (i.e., arms) [16]. However, these results do not apply to the online settings with *unknown* distributions considered in this paper.

III. System Model and Problem Formulation

In this section, we define contextual bandits with probing (CBwP) as a novel extension of the classic contextual bandits model [4]. To make it concrete, we use the joint AP probing and selection problem as an example when presenting the model. Our formulation applies to a large class of sequential decision problems that involve joint probing and selection under uncertainty. We further derive some important properties of CBwP.

A. Contextual Bandits with Probing (CBwP)

We consider a set of APs connected to a high-speed backhaul that collaboratively serve a set of mobile clients. AP collaboration helps boost wireless performance in both indoor and outdoor environments and is especially useful for directional mmWave communications that are susceptible to blockage [10]. For simplicity, we assume that the beamforming process determines the best beam (i.e., highest SNR) from AP to the client. Hence, we do not distinguish AP selection from beam selection. Our framework readily applies to the more general setting of joint AP and beam selection.

To simplify the problem, we focus on the single client setting in this work. Let X be a set of contexts that correspond to the location (or a rough estimate of it) of a moving client. In general, X can be either discrete or continuous. Let A be a discrete set of arms that correspond to the set of APs, and $N \triangleq |A|$. We consider a fixed time horizon T that is known to the decision maker. In each time step t, the decision maker first receives a context $x_t \in X$ and then plays an arm $a_t \in A$ and receives a reward $\phi(a_t|x_t) \in [0,1]$, which is $i.i.d.$ sampled from an *unknown* distribution, $\Phi(a_t|x_t)$, that depends on both the context x_t and the arm a_t. We assume that the expected value of $\Phi(a|x)$ exists for any context-arm pair (x,a) and denote it by $\mu(a|x)$. The sequence of context $(x_t)_{t\in\mathbb{N}}$ is assumed to be external to the decision making process. In the AP selection problem, the reward corresponds to the data rate that a client at a certain location can receive from an AP.

In the classic contextual bandit problem, the instantaneous reward of an arm is revealed only when it is played, and the decision maker receives no side observations. In contrast, we consider a more general setting where after receiving the context, the decision maker can first probe a subset of $K < N$ arms and observe their rewards, and then pick an arm to play (which may be different from the set of probed arms). In general, probing an arm reduces the uncertainty about the arm. We assume that the probing period (for K arms) is short enough so that if an arm a is probed with $\phi(a|x_t)$ observed, then the same $\phi(a|x_t)$ is the reward obtained if arm a is played in t. However, probing does reduce packet transmission time; hence, we require K to be relatively small. The problem of choosing a proper K either statically or dynamically is left to our future work. We further assume that the probing results are independent across arms. That is, $\phi(a|x_t)$ is independent of other arms probed in t or before t. Extension to correlated arms is left to future work.

Let $G_{\pi_t}(x_t)$ denote the *expected* reward in time step t under a (time-varying) joint probing and play policy π_t, where the expectation is over the randomness of observations in time step t. Similar to the classic contextual bandit model, our goal is to maximize the (expected) cumulative reward $\mathbb{E}[\sum_{t=1}^{T} G_{\pi_t}(x_t)]$. As we discuss below, when the reward distribution $\Phi(a|x)$ is known a prior for each context-arm pair (x,a), the single stage problem at each time step can be modeled as a Markov decision process with an optimal policy. Let $G^\star(x_t)$ denote the expected reward under x_t when the optimal *offline* policy

is adopted in each time step. Define the total regret as follows:

$$R(T) = \sum_{t=1}^{T} (G^\star(x_t) - G_{\pi_t}(x_t)) \quad (1)$$

The goal of maximizing the expected cumulative reward then converts to minimizing $\mathbb{E}(R(T))$.

Similar to [4], we assume that the context set X is associated with a distance metric $\mathcal{D}$ such that $\mu(a|x)$ satisfies the following Lipschitz condition:

$$|\mu(a|x) - \mu(a|x')| \leq \mathcal{D}(x, x'), \forall a \in A, x, x' \in X \quad (2)$$

Without loss of generality, we assume that $\mathcal{D}(\cdot, \cdot) \leq 1$. This condition helps us capture the similarity between the context-arm pairs. In the joint AP probing and selection problem, $\mathcal{D}$ is defined as the Euclidean distance between locations.

B. Offline Problem as an MDP

We first consider the offline setting where the reward distributions are known to the decision maker a prior. We show that the joint probing and play problem in each time step can itself be modeled as a Markov decision process (MDP). We further derive important properties of the MDP. Due to the space limitation, we omit all the proofs. The reader is referred to [17] for the missing details.

Consider any time step with a context x. To simplify the notation, we omit the time step subscript in this section. At each probing step $i \in \{0, 1, ...K\}$, the decision maker observes the current state $s_i \triangleq (a_1, \ldots, a_i, \phi(a_1|x), \ldots, \phi(a_i|x))$ and then chooses the next arm a_{i+1} to probe, where a_j is the arm probed in round j and $\phi(a_j|x)$ is the observed reward of arm a_j under context x. We define $s_0 \triangleq \emptyset$. Further, the decision maker can decide at any round $i \leq K$ to stop probing and pick an arm to play according to the probing result, and receives the reward of the played arm. We observe that more information always helps in our problem, thus it never hurts to wait until round K to choose an arm to play. Let S denote the set of all possible states and $\mathcal{P}(A)$ the set of distributions over A. The joint problem can be solved using a pair of policies: a probing policy $\pi_1 : S \to \mathcal{P}(A)$ that maps an arbitrary probing history to the next arm to probe and a play policy $\pi_2 : S \to \mathcal{P}(A)$ that chooses an arm to play according to the probing result. Let $\pi = (\pi_1, \pi_2)$ denote a joint probing and play policy.

We first observe that in the offline setting, there is a simple deterministic play policy that is optimal. Let $g_{\pi_2}(s_i)$ denote the expected reward that can be obtained from playing an arm using policy π_2 given the probing result s_i after i rounds. We have

$$g_{\pi_2}(s_i) = \sum_{j=1}^{i} \pi_2(a_j|s_i)\phi(a_j|x) + \sum_{b_j \in A \setminus \{a_1, \ldots, a_i\}} \pi_2(b_j|s_i)\mu(b_j|x)$$

where $\pi_2(a|s)$ denotes the probability of playing arm a given the probing result s. For any arm a, let $v(a|x, s_i) = \phi(a|x)$ if $a \in \{a_1, ..., a_i\}$ and $v(a|x, s_i) = \mu(a|x)$ otherwise. Then we observe that the deterministic policy that always plays an arm with maximum $v(a|x, s_i)$ is optimal and obtains the following optimal reward:

$$g^\star(s_i) = \max\Big\{ \max_{a \in \{a_1, \ldots, a_i\}} \phi(a|x_t), \max_{b \in A \setminus \{a_1 \ldots, a_i\}} \mu(b) \Big\} \quad (3)$$

We summarize this observation as a lemma:

Lemma 1: Given any context x and state s_i, the deterministic policy that plays an arm with maximum $v(a|x, s_i)$ is optimal.

We then consider the problem of finding an optimal probing policy. For any given play policy π_2, the probing problem can be formulated as a finite-horizon MDP $M = (S, A, rwd, tr, K)$, where S is a set of states defined above, A is set of actions that correspond to the set of arms. The reward function $rwd : S \times A \times S \to \mathbb{R}$ is defined as $rwd(s_i, a_{i+1}, s_{i+1}) = g_{\pi_2}(s_{i+1}) - g_{\pi_2}(s_i)$ for $i < K$ and $rwd(s_i, a_{i+1}, s_{i+1}) = 0$ otherwise. The transition dynamics $tr(s_{i+1}|s_i, a_{i+1})$ gives the probability of reaching state s_{i+1} given the current state s_i and action a_{i+1}, which can be derived from $\Pr(s_{i+1} = (s_i, a_{i+1}, \phi(a_{i+1}|x))|s_i, a_{i+1}) = \Pr(\Phi(a_{i+1}|x) = \phi(a_{i+1}|x))$.

We consider the standard objective of maximizing the expected cumulative reward for the MDP. Given the way the reward function is defined, this can be represented as $G_\pi \triangleq \mathbb{E}_{\pi,\Phi}[\sum_{i=1}^{K-1} rwd(s_i, a_{i+1}, s_{i+1})] = \mathbb{E}_{\pi_1,\Phi}[g_{\pi_2}(s_K)]$. Thus, to find the optimal π, it suffices to adopt an optimal play policy $\pi_2^\star$ (such as the deterministic policy defined above) and solve the MDP to find the optimal probing policy $\pi_1^\star$. Let $\pi^\star = (\pi_1^\star, \pi_2^\star)$ denote the optimal joint (offline) policy.

The MDP M defined above uses the complete history of the probing results as the state. We then show that assuming an optimal play policy is adopted, it suffices to keep the set of arms probed and the maximum reward observed. This allows us to obtain a smaller MDP without loss of optimality. To show this, given any state $s_i = (a_1, ..., a_i, \phi(a_1|x), ..., \phi(a_i|x))$, we derive a new state $\overline{s}_i \triangleq (a_1, ..., a_i, \max(\phi(a_1|x), ..., \phi(a_i|x)))$. Let $\overline{S}$ denote the set of states $\overline{s}$. We further say that s is similar to $\overline{s}$ (denoted by $s \sim \overline{s}$) if the latter can be derived from the former. We then define a new MDP $M' = (\overline{S}, A, rwd', tr', K)$, where $rwd'(\overline{s}_i, a_{i+1}, \overline{s}_{i+1}) = g^\star(s_{i+1}) - g^\star(s_i)$ for any s_i and s_{i+1} such that $s_i \sim \overline{s}_i$ and $s_{i+1} \sim \overline{s}_{i+1}$. Note that the reward function is well defined as $g^\star(s_i)$ only depends on the maximum probed value in s_i (see Equation (3)). Further, the new transition dynamics tr' can be derived from the following observation:

$$\Pr(\overline{s}_{i+1}|\overline{s}_i, a_{i+1}) = \sum_{s_{i+1} \sim \overline{s}_{i+1}} \Pr(s_{i+1}|s_i, a_{i+1}), \forall s_i \sim \overline{s}_i.$$

We then show that M and M' have the same optimal value. Let $Q_M^\star(s_i, a_{i+1}) \triangleq \max_\pi \mathbb{E}_{\pi,\Phi}[\sum_{j=i}^{K-1} rwd(s_j, a_{j+1}, s_{j+1})|s_i, a_{i+1}]$ denote the optimal state-action value function of M for any state s_i and action a_{i+1}. Assuming $Q_M^\star(s_K, a_{K+1}) = 0$, $Q_M^\star$ satisfies the Bellman optimality equation:

$$Q_M^\star(s_i, a_{i+1}) = \sum_{s_{i+1} \in S} \Pr(s_{i+1}|s_i, a_{i+1}) \big[rwd(s_i, a_{i+1}, s_{i+1})$$

$$+ \max_{a' \in A} Q^{\star}_{M}(s_{i+1}, a')\big]$$

$Q^{\star}_{M'}(\overline{s}_i, a_{i+1})$ is defined analogously. We then have the following result, which can be derived using the Bellman optimality equation and mathematical induction:

Lemma 2: $Q^{\star}_{M}(s_i, a_{i+1}) = Q^{\star}_{M'}(\overline{s}_i, a_{i+1}), \quad \forall s_i \sim \overline{s}_i.$

C. Offline Problem with Bernoulli Rewards

When $\Phi(a|x)$ follows a Bernoulli distribution (fully defined by $\mu(a|x)$) for any context-arm pair (x, a). there is a simple non-adaptive probing policy that is optimal.

Lemma 3: Consider the following non-adaptive probing policy for arms with Bernoulli rewards: given any context x, find the $K+1$ arms with the largest $\mu(a|x)$ among all the arms, and then probe any K of them. This policy together with the deterministic play policy given in Lemma 1 gives an optimal joint policy to the offline problem.

IV. The CBwP Algorithm

Lemma 3 indicates that in the offline setting, greedy probing plus greedy play is optimal for Bernoulli rewards. However, this approach cannot be directly applied to the online setting as $\Phi(a|x)$'s are unknown, which involves a fundamental exploration vs. exploitation tradeoff. In this section, we consider the online setting and design an algorithm for the CBwP problem. We further derive its regret for the special case with Bernoulli rewards.

A. Algorithm Description

Our algorithm is based on the contextual zooming algorithm in [4], which adaptively partitions the similarity space to exploit the Lipschitz condition (Equation (2)). As we consider a finite set of arms in our problem, we apply adaptive partitioning to the context space only. The main contribution of our work is extending the contextual zooming technique to the joint probing and play setting, which brings new challenges in both algorithm design and analysis.

The algorithm (see Algorithm 1) maintains a finite set $\mathcal{A}_a$ of active balls for each arm a. We require that the balls in $\mathcal{A}_a$ collectively cover the similarity space $(X, \mathcal{D})$. Initially $\mathcal{A}_a$ contains a single ball of radius 1. A ball is activated once it is added to $\mathcal{A}_a$ and remains active. These balls correspond to a partition of the context space from the arm a's perspective.

In each time round t, a context x_t is revealed, and the algorithm selects up to K arms $\{a_1, a_2, \ldots, a_K\}$ to probe according to the "probing rule". In each probing step after observing $\phi(a_i|x_t)$, the algorithm may activate a new ball according to the "activation rule". The probing stage terminates if either K arms have been probed or $\phi(a_i|x_t) = 1$ for some i. The algorithm then selects an arm b to play according to the "playing rule", and may activate a new ball according to the "activation rule".

We then specify the three rules used in the algorithm. Both the probing and play rules are inspired by Lemma 3. Since the true distributions of rewards are unknown, the algorithm picks arms according to their estimated rewards together with a confidence term. Let $r(B)$ denote the radius of a ball B. The *confidence radius* of B at time round t is defined as:

$$\mathrm{conf}_t(B) \triangleq 4\sqrt{\frac{\log T}{1 + n_t(B)}} \tag{4}$$

where $n_t(B)$ is the number of times B has been selected from $t = 1$ to t. Let $\mathrm{re}_t(B)$ denote the total reward from all rounds up to $t-1$ in which B has been selected, and $v_t(B) \triangleq \frac{\mathrm{re}_t(B)}{\max(1, n_t(B))}$ the average reward from B. The *pre-index* of B is defined as

$$I^{\mathrm{pre}}_t(B) = v_t(B) + r(B) + \mathrm{conf}_t(B) \tag{5}$$

The *index* of $B \in \mathcal{A}_a$ is obtained by taking a minimum over all active balls of AP a:

$$I_t(B) \triangleq r(B) + \min_{B' \in \mathcal{A}_a} (I^{\mathrm{pre}}_t(B') + \mathcal{D}(B, B')), \forall B \in \mathcal{A}_a \tag{6}$$

where $\mathcal{D}(B_a, B'_a)$ is the distance between the centers of the two balls.

In each round t and for each AP a, let $\mathcal{B}_a \subseteq \mathcal{A}_a$ be the set of active balls that contains x_t and has the minimum radius. Let B^{sel}_a be an arbitrary ball in $\mathcal{B}_a$ with maximal index. We then state the three rules as follows:

- **probing rule**: At each probing step i in time round t, the algorithm probes an arm a_i with the maximal index $I_t(B^{\mathrm{sel}}_a)$ (break ties arbitrarily) among the set of unprobed arms and gets observation $\phi(a_i|x_t)$. The probing stage ends if K arms have been probed or $\phi(a_i|x_t) = 1$ for some i.
- **playing rule**: In each time round t and after the probing stage is done, let $v(a) = \phi(a|x_t)$ if a has been probed and $v(a) = I_t(B^{\mathrm{sel}}_a)$ otherwise. If $\phi(a_i|x_t) = 1$ for some i in the probing stage, play a_i. Otherwise, play an arm b with maximal $v(b)$.
- **activation rule**: If arm a is probed or played in time round t, the algorithm updates $n_t(B^{\mathrm{sel}}_a)$ and $\mathrm{conf}_t(B_a)$. Further, a new ball with center x_t and radius $\frac{1}{2}r(B^{\mathrm{sel}}_a)$ is activated if $\mathrm{conf}_t(B^{\mathrm{sel}}_a) \leq r(B^{\mathrm{sel}}_a)$. B^{sel}_a is called the parent ball of this newly activated ball.

Remark 1: We note that the index $I_t(B)$ defined above includes both the average reward $v_t(B)$ and a confidence radius, similar to the classic upper confidence bound (UCB) based approaches [7]. Further, exploration is included in both probing and play stages. One may wonder if this is necessary since probing provides free observations, which may remove the necessity of exploration. However, as we show in our simulations, replacing $I_t(B)$ by $v_t(B)$ (so that no exploration is used) leads to suboptimal decisions. Intuitively, this is because although probing reduces uncertainty, it does not completely remove it for a small K. Thus, it is crucial to judiciously utilize the limited probing resource.

B. Theoretical Analysis for CBwP with Bernoulli rewards

In this section, we analyze the regret of Algorithm 1 in the special case when the rewards of arms follow Bernoulli distributions. Consider the optimal policy $\pi^{\star}$ defined in Lemma

Algorithm 1 Contextual zooming for joint probing and play

Input: A: a set of N arms; $(X, \mathcal{D})$: a similarity space of diameter ≤ 1; T: time horizon

for each AP a **do**
  $B \leftarrow B(x, 1)$ //center $x \in X$ is arbitrary
  $\mathcal{A}_a \leftarrow \{B\}$, $n(B) \leftarrow 0$, $\text{re}(B) \leftarrow 0$
for each round $t = 1, 2, \ldots, T$ **do**
  Input context x_t
  for each AP a **do**
    $\mathcal{B}_a \leftarrow \{B \in \mathcal{A}_a : r(B) = \min_{B' \in \mathcal{A}_a, x_t \in B'} r(B')\}$
    $B_a^{\text{sel}} \leftarrow \operatorname{argmax}_{B \in \mathcal{B}_a} I_t(B)$
    $v(a) \leftarrow I_t(B_a^{\text{sel}})$
  for $i = 1$ to K **do**
    $a_i \leftarrow \operatorname{argmax}_{a_i' \in A \setminus \{a_1, a_2, \ldots, a_{i-1}\}} I_t(B_{a_i'}^{\text{sel}})$ //Probing
    Probe a_i, get the observation $\phi(a_i|x_t)$
    $v(a_i) \leftarrow \phi(a_i|x_t)$
    ACTIVATION($\phi(a_i|x_t), B_{a_i}^{\text{sel}}, x_t$)
    if $\phi(a_i|x_t) = 1$ **then**
      Break
  if $\phi(a_i|x_t) = 1$ for some i **then** //Playing
    Play arm a_i, get the reward 1
  else
    $b \leftarrow \operatorname{argmax}_{b \in A} v(b)$
    Play arm b, get the reward $\phi(b|x_t)$
    ACTIVATION($\phi(b|x_t), B_b^{\text{sel}}, x_t$)
function ACTIVATION($\phi(a|x_t), B_a, x_t$):
  $n(B_a) \leftarrow n(B_a) + 1$
  $\text{re}(B_a) \leftarrow \text{re}(B_a) + \phi(a|x_t)$
  if $\text{conf}(B_a) \leq \text{radius}(B_a)$ **then** //Activation
    $B' \leftarrow B(x_t, \frac{1}{2}\text{radius}(B_a))$
    $\mathcal{A}_a \leftarrow \mathcal{A}_a \bigcup \{B'\}$
    $n(B') \leftarrow 0$, $\text{re}(B') \leftarrow 0$

3. For Bernoulli bandits, we have $G_{\pi_t}(x_t) = 1 - \prod_{i=1}^{K+1}(1 - \mu(a_{it}|x_t))$ for any π_t that makes use of probing results, where a_{it} is an arm probed or played in round t chosen by π_t. Using mathematical induction, we can derive the following bound for the regret in each round.

Lemma 4: For any round t, we have

$$\begin{aligned}\Delta(G_{\pi_t}|x_t)) \triangleq & G_{\pi^\star}(x_t) - G_{\pi_t}(x_t) \\ \leq & (1 - \min_{i \in \{1,2,\ldots,K\}} \mu(a_{it}|x_t))^K \sum_{j=1}^{K+1} \Delta(a_{jt}|x_t).\end{aligned}$$

To bound the total regret, the main idea is to show that with high probability, (1) $\Delta(a_{it}|x_t)$ is bounded by $r(B_{a_{it}}^{\text{sel}})$ times a constant for any a_{it} and x_t, and (2) the total number of balls associated with any given arm that have radius r and have been activated throughout the execution of the algorithm is bounded by N_r, the r-packing number [4]. We then have the following main result. The detailed proof can be found in our online technical report [17].

Theorem 1: The total expected regret of Algorithm 1 is bounded as follows:

$$\begin{aligned}\mathbb{E}(R(T)) \leq & (K+1)(14 r_0 HT + 2) \\ & + 224NH \left(\sum_{r = 2^{-j} : r_0 \leq r \leq 1} r^{-1} N_r \right) \log T\end{aligned}$$

where $H \triangleq (1 - \min_{a \in A, x \in X} \mu(a|x))^K$.

Note that the bound in the theorem can be tightened by taking inf on all $r_0 \in (0, 1)$.

V. Evaluation

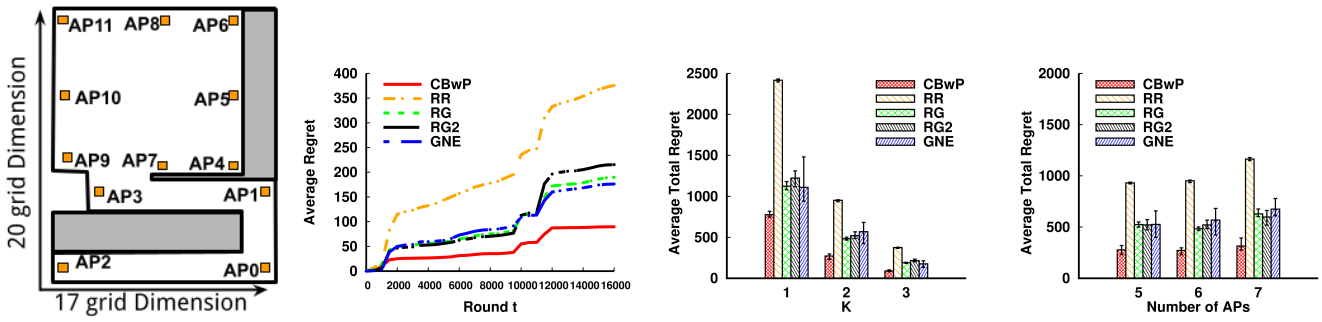
In this section, we present the evaluation results of CBwP by comparing it with four baselines: (a) RR (random probing and random play): randomly probing K arms and then randomly playing an arm (if none of the probed arms gives a reward of 1); (b) RG (random probing and greedy play with exploration): randomly probing K arms and then playing the arm a with maximum $v(a)$ (as defined in the playing rule in Section IV); (c) RG2 (random probing and greedy play without exploration): randomly probing K arms and then playing the arm a with maximum $v'(a)$, where $v'(a) = \phi(a|x_t)$ if a has been probed and $v'(a) = v_t(B_a^{\text{sel}})$ otherwise; (d) GNE (greedy probing and play without exploration): this is a variant of Algorithm 1 where we replace the index of a ball $I_t(B)$ with its average reward $v_t(B)$ in both the probing the playing rules.

A. Evaluation Settings

In order to evaluate the system, we collect the channel traces from the real testbeds with 802.11ad routers and laptops, and a commercial mmWave channel simulator (Remcom Insite [18]). We collected SNR traces in a student hall (Scenario 1) with real testbeds at 250 different locations at the granularity of 0.8m. 12 Airfide [19] 802.11ad APs are deployed in the student hall and each of them is equipped with 8 phased array antennas with 64 sectors. We use the Acer TravelMate-648 laptops with a single phased array and 36 sectors as the clients. We modified the open-source driver of both devices to extract the SNR and beamforming information.

After getting the signal strength channel traces from our testbed and the Remcom channel simulator, we use MCS-SNR table from the 802.11ad [20] to map the SNR to get the average throughput of the link (as in [21]) based on the best Tx/Rx sector pair of the link found using the beamforming process. We normalize the average throughput of each AP a as $\mu(a|x)$ for each context x (location of a client). We consider Bernoulli distribution with probability $\mu(a|x)$ to get reward 1 for each time round.

For the mobility traces, we select from 15 to 80 clients randomly located where each client follows a specific walking pattern. We set the walking pattern of the clients by observing the typical walking behaviors in each room. We assume the walking speed as $1m/s$ for all clients and for each time slot, the client will locate at one of the grids where the channel is measured by our testbed or the channel simulator as described before. In the simulations, we choose 10 clients' traces where each of the traces has 80 steps and each step includes 20 time rounds. For computing the distance $\mathcal{D}(x, x')$ between two arbitrary locations x and x', we let $\mathcal{D}(x, x') = E(x, x')/\text{Dia}$



(a) (b) (c) (d)

Fig. 1: (a) Scenario: Student Hall; (b)-(d) Comparing CBwP with baselines on average regret: (b) $K = 3, N = 6$; (c) $N = 6$; (d) $K = 2$.

where $E(x, x')$ is the Euclidean distance between x and x' and Dia is the longest diagonal line length.

We note that although we consider an indoor environment in the evaluation, our approach can be readily applied to outdoor settings as well, e.g., mmWave based vehicular networks [10].

B. Results

We evaluate all the algorithms with different K and number of APs N. For each N, we randomly pick 5 different sets of APs with size N. In each setting, we run all the algorithms 10 times (with the same random seeds) and take the average.

Fig. 1b shows how the average total expected regret changes with time. Compared with the baselines, CBwP's regret increases more slowly and after time round 1,500, the average regret of CBwP's is always lower than the others. In the figure, there are two jumps around time rounds 9,500 and 11,000, respectively, which correspond to the starting points of two new clients who entered the room from the locations less explored in previous time rounds, e.g., the bottom area where most APs are blocked.

Fig. 1c shows the average total expected regret of all the five algorithms under different K with $N = 6$. We observe that CBwP outperforms all the baselines irrespective of K. In addition, with the value of K increasing, the average expected regret gradually declines in all the algorithms. This is expected as a larger K provides a higher chance of finding a good arm to play. Fig. 1d shows the average total expected regret of the five algorithms for different numbers of APs with $K = 2$. CBwP again outperforms all the baselines. Further, with AP increasing, the average expected regret gradually rises for the baselines, indicating their difficulty of scaling to more APs. In contrast, the performance of CBwP is stable.

VI. Conclusion

In this paper, we consider the problem that APs cooperatively serve a mobile client with unknown date rates. We propose contextual bandit with probing (CBwP) as a novel bandit learning framework that incorporates joint probing and play to solve this problem. We derive structural properties of the optimal offline solution and an efficient online learning algorithm to CBwP. We further establish the regret bound of CBwP for links with Bernoulli data rates. Our CBwP model is a novel extension to the classic contextual bandit model and can potentially be applied to a large class of sequential decision-making problems that involve joint probing and play under uncertainty.

Acknowledgment

This work was supported in part by NSF grant CNS-1816943.

References

[1] T. Stahlbuhk, B. Shrader, and E. Modiano, "Learning algorithms for scheduling in wireless networks with unknown channel statistics," *Ad Hoc Networks*, vol. 85, no. 2019, pp. 131–144, 2019.

[2] F. Li, J. Liu, , and B. Ji, "Combinatorial sleeping bandits with fairness constraints," in *IEEE INFOCOM*, 2019.

[3] H. Gupta, A. Eryilmaz, and R. Srikant, "Link rate selection using constrained thompson sampling," in *IEEE INFOCOM*, 2019.

[4] A. Slivkins, "Contextual bandits with similarity information," in *COLT*, 2011.

[5] A. Bhaskara, S. Gollapudi, K. Kollias, and K. Munagala, "Adaptive probing policies for shortest path routing," in *NeurIPS*, 2020.

[6] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[8] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[9] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[10] A. Asadi, S. Müller, G. H. Sim, A. Klein, and M. Hollick, "FML: Fast Machine Learning for 5G mmWave Vehicular Communications," in *IEEE INFOCOM*, 2018.

[11] K. Munagala, S. Babu, R. Motwani, and J. Widom, "The pipelined set cover problem," in *International Conference on Database Theory*, pp. 83–98, Springer, 2005.

[12] Z. Liu, S. Parthasarathy, A. Ranganathan, and H. Yang, "Near-optimal algorithms for shared filter evaluation in data stream systems," in *ACM SIGMOD*, 2008.

[13] A. Deshpande, L. Hellerstein, and D. Kletenik, "Approximation algorithms for stochastic submodular set cover with applications to boolean function evaluation and min-knapsack," *ACM Transactions on Algorithms (TALG)*, vol. 12, no. 3, pp. 1–28, 2016.

[14] S. Guha, K. Munagala, and S. Sarkar, "Optimizing transmission rate in wireless channels using adaptive probes," in *ACM Sigmetrics*, 2006.

[15] A. Goel, S. Guha, and K. Munagala, "Asking the right questions: Model-driven optimization using probes," in *PODS*, 2006.

[16] C. Attias, R. Krauthgamer, R. Levi, and Y. Shaposhnik, "Stochastic selection problems with testing," *Available at SSRN 3076956*, 2017.

[17] T. Xu, D. Zhang, P. H. Pathak, and Z. Zheng, "Joint AP Probing and Scheduling: A Contextual Bandit Approach." Technical Report, available online at https://arxiv.org/abs/2108.03297.

[18] "Remcom Wireless InSite 3D Wireless Propagation Software." https://www.remcom.com/wireless-insite-em-propagation-software.

[19] "Airfide afn2200." https://airfidenet.com/.

[20] IEEE P802.11adTM/D4.0, "Part 11: Wlan mac/phy enhancements for very high throughput in the 60 ghz band," 2012.

[21] D. Zhang, M. Garude, and P. H. Pathak, "Mmchoir: Exploiting joint transmissions for reliable 60ghz mmwave wlans," in *ACM Mobihoc*, 2018.