Online Bayesian Learning for Rate Selection in Millimeter Wave Cognitive Radio Networks

Muhammad Anjum Qureshi, Cem Tekin
Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey
{qureshi, cemtekin}@ee.bilkent.edu.tr

***Abstract*—We consider the problem of dynamic rate selection in a *cognitive radio network* (CRN) over the *millimeter wave* (mmWave) spectrum. Specifically, we focus on the scenario when the transmit power is time varying as motivated by the following applications: i) an energy harvesting CRN, in which the system solely relies on the harvested energy source, and ii) an underlay CRN, in which a *secondary user* (SU) restricts its transmission power based on a dynamically changing *interference temperature limit* (ITL) such that the *primary user* (PU) remains unharmed. Since the channel quality fluctuates very rapidly in mmWave networks and costly *channel state information* (CSI) is not that useful, we consider rate adaptation over an mmWave channel as an online stochastic optimization problem, and propose a *Thompson Sampling* (TS) based Bayesian method. Our method utilizes the unimodality and monotonicity of the throughput with respect to rates and transmit powers and achieves logarithmic in time regret with a leading term that is independent of the number of available rates. Our regret bound holds for any sequence of transmits powers and captures the dependence of the regret on the arrival pattern. We also show via simulations that the performance of the proposed algorithm is superior than the state-of-the-art algorithms, especially when the arrivals are favorable.**

***Index Terms*—Cognitive radio networks, mmWave, dynamic rate selection, Thompson sampling, contextual unimodal bandits.**

I. Introduction

The immense expansion of the number of wireless devices and mobile services has recently enforced Federal Communications Commission (FCC) to open up the portion of vast millimeter Wave (mmWave) spectrum band that spans between 30GHz to 300GHz for wireless communications [1]. However, making this spectrum band a feasible resource for the next-generation wireless services requires dealing with its unique and highly dynamic characteristics such as high path loss, signal attenuation and atmospheric absorption [2], [3]. In particular, highly dynamic nature and environmental dependency of mmWave makes channel state information (CSI) based resource allocation algorithms impractical in this setting [4].

On the one hand, statistical characteristics of mmWave communication motivates learning theory based solutions to perform resource allocation tasks [4]–[9]. On the other hand, AI-enabled cognitive radio networks (CRN) are conceived to further enhance the spectral efficiency of the mmWave band [10]–[12]. Moreover, energy harvesting based solutions are becoming ubiquitous for many self-sustainable real-world systems, where energy is continually proliferated from nature or man-made phenomena instead of conventional battery-powered generation [13], [14]. All of these motivates us to consider the problem of dynamic rate selection in a CRN that operates over the mmWave band with a time-varying transmit power.

In these networks, the transmit power usually needs to be adjusted based on exogenous events. For instance, in the spectrum underlay paradigm [15], secondary users (SUs), are capable of sensing the spectrum and adapt their power so that the interference to primary users (PUs) remains below a threshold. The term interference temperature limit (ITL) sets a pre-defined threshold, which needs to be satisfied as long as the SU is using the specific frequency band. ITL is dependent on numerous factors including the location of the SU and the selected spectrum frequency [16]. As another example, in an energy harvesting CRN without any explicit battery or super-capacitor, the harvested energy is used by the system instantly without any storage. In the considered scenario, the transmit power of the SU is dependent on the current harvested energy.

Rate adaptation (RA) over a wireless communication system is a fundamental mechanism that allows the transmitter to adapt the modulation and coding scheme to the channel conditions, where the target is to maximize the overall throughput which is defined as the product of the rate and the packet success probability over that rate [4], [6]–[8]. The packet transmission outcome is random and the packet success probabilities are not known a priori to the transmitter. These probabilities depend on the transmission power, and they need to be learned via interacting with the environment. We assume that the only feedback available after a transmission is the ACK/NAK flag. The transmitter has to learn the best rate via utilizing this feedback and taking into account its input parameters, which motivates us to develop a new online adaptive allocation strategy to learn faster.

We rigorously formulate the aforementioned problem as a contextual Bayesian unimodal multi-armed bandit (MAB). We refer to each modulation and coding scheme (MCS) as an arm and to the transmit power as a side information (context). The user (learner) selects an arm after observing the context in each round. The expected reward is defined as the throughput and consistent with real-world observations [7], [17] it is assumed to exhibit unimodal structure under different MCSs and a

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 116E229 and by BAGEP Award of the Science Academy (2019). (*Corresponding author: Muhammad Anjum Qureshi.*)

monotone structure under different transmit powers.

Earlier works on rate adaptation formalize the problem as a non-contextual MAB. For example, [6] and [7] propose MAB algorithms based on Kullback-Leibler upper confidence bound (KL-UCB) indices that learn the optimal rate via utilizing the unimodal structure of the expected reward over the rates. Authors in [18] propose a contextual learning algorithm based on KL-UCB, which exploits unimodality of the expected reward both in the arms and the contexts. On the other hand, [4] exploits rate unimodality by using a variant of Thompson sampling and shows that the regret grows logarithmically over time. It is shown in [4] and [19] via numerical experiments that Thompson sampling outperforms the frequentist approach based on KL-UCB indices. Similar to these works, we propose a Bayesian approach that exploits the unimodality of the expected reward over rates and monotonicity of the expected reward over transmit powers to learn fast.

Exogeneity of the contexts makes regret analysis significantly different than non-contextual versions of Thompson sampling that use the structure in rewards [4], [8], [19]. It is shown in [4] that *modified Thompson sampling* (MTS) can achieve logarithmic regret by keeping independent priors for arms and by decoupling rate from success probability. However, the important structural property of rate unimodality is not exploited in MTS. It is shown in [19] that under a unimodal assumption on the expected reward function, it is also possible to achieve logarithmic regret. The proposed *unimodal Thompson sampling* (UTS) keeps independent priors and exploits the arm unimodality like [20]. However, this algorithm is for general expected rewards and does not decouple the rate. A similar algorithm with detailed analysis for rank one bandits is proposed in [21], which explicitly calculates the constants in the regret. In [8], authors propose *constrained Thompson sampling* (CoTS) that exploits the structure more efficiently than MTS by assuming that the success probability is monotonic over the rate. In contrast to all these prior works, which do not consider contexts, our algorithm exploits the unimodal structure over arms, the contextual information and the monotone structure of the reward over the contexts.

Our main contributions are summarized as follows:

- We consider the problem of rate selection under time-varying transmit power over an mmWave channel and formalize the problem as a contextual MAB.
- We propose a Bayesian learning algorithm called DRS-TS which exploits the structure among rates as well as among transmit powers. We prove that the regret of DRS-TS scales logarithmically in time and the leading term in the regret is independent of the number of rates. To the best of our knowledge, this is the first work that analyses Thompson sampling in a contextual unimodal MAB.
- We compare DRS-TS with other state-of-the-art learning algorithms and show that it significantly outperforms its competitors via numerical experiments.

II. Problem Formulation

We consider a single transmitter and receiver based wireless link, where the transmitter (i) is powered from an energy harvested source in case of the energy harvesting CRN or (ii) obeys an ITL in case of the underlay CRN when selecting its transmit power. The transmitter can transmit by using one of the K available transmission rates by choosing the corresponding MCS and one of the P possible transmit powers.[1] Let $\mathcal{K} = [K]$ denote the set of MCS and r_i, $i \in \mathcal{K}$, denote the rate that corresponds to the ith MCS. The set of rates $\mathcal{R} = \{r_1, \ldots, r_K\}$ and the set of transmit powers $\mathcal{P} = \{p_1, \ldots, p_P\}$ are ordered such that $r_1 < \ldots < r_K$ and $p_1 < \ldots < p_P$. For $p \in \mathcal{P}$ and $r \in \mathcal{R}$, j_p and i_r represent the indices of transmit power p and rate r, i.e., $p_{j_p} = p$ and $r_{i_r} = r$.

It is well known that dynamic rate selection over rapidly varying wireless channels can be modeled as a MAB problem. Moreover, it is shown that this formulation is asymptotically equivalent to maximizing the number of packets successfully transmitted over a given time horizon [4]–[9]. Therefore, in the rest of the paper we describe and analyze the equivalent MAB formulation.

In the MAB formulation, the SU makes decisions sequentially over rounds indexed by $t \in [T]$, where T represents the time horizon. At the beginning of round t, the SU receives the transmit power it should use in that round, denoted by $p(t)$ with index $j(t)$, i.e., $p(t) = p_{j(t)}$. In case of the energy harvesting CRN, this comes from the power control algorithm of the energy harvesting device [14], while in case of the underlay CRN, this comes from interference temperature measurements [16]. Then, the SU chooses a modulation and coding scheme from $\mathcal{K}$ with corresponding rate $r(t)$ and transmits with power $p(t)$ and rate $r(t)$. At the end of the round it receives ACK/NAK feedback $x_{p(t),r(t)}(t)$ indicating whether the transmission was successful or not, and collects as throughput the (normalized) reward $g_{p(t),r(t)}(t) = (r(t)/r_K)x_{p(t),r(t)}(t)$. Here, $x_{p,r}(t)$ is a Bernoulli random variable with expected value $\psi(p,r)$ that takes value 1 if the transmission given power-rate pair (p,r) in round t is successful and value 0 otherwise. We call $\psi(p,r)$ the transmission success probability and $\mu(p,r) = (r/r_K) \cdot \psi(p,r)$ the normalized throughput associated with power-rate pair (p,r). We also call transmit powers *contexts* and rates *arms* of the MAB.

In case of the energy harvesting CRN, we consider no embedded energy supply so that the CRN solely relies on the harvested energy. The dynamic power output from an energy harvesting source (e.g., RF energy source, solar cell, wind turbine) is directly used by the load [14]. In case of the underlay CRN, the SU uses its cognitive capabilities to determine the transmit power such that the dynamic ITL is satisfied [24]. For a single dynamic SU, ITL is dependent on the current location of the SU [25], and assuming known primary location, the SU adjusts its transmit power to satisfy the current ITL. For cooperative multi-player homogeneous CRN in which the SUs are assigned channels in round robin

[1] Similar to [22] and [23], we adopt a discrete power set.

manner, ITL which is dependent on the channel frequency, is calculated beforehand for each frequency channel [26], and the SU adjusts its transmit power to satisfy ITL for the currently assigned frequency channel.

The optimal rate given a transmit power p is denoted by $r_p^* = \text{argmax}_{r\in\mathcal{R}}\, \mu(p,r)$ and its index is given as i_p^*, i.e., $r_p^* = r_{i_p^*}$. Without loss of generality we assume that r_p^* is unique. The suboptimality gap of rate r given transmit power p is defined as $\Delta(p,r) = \mu(p,r_p^*) - \mu(p,r)$. The set of neighbors of rate r_i is given as $\mathcal{N}(r_i)$. We have $\mathcal{N}(r_i) = \{r_{i-1}, r_{i+1}\}$ for $i \in \{2,\dots,K-1\}$, $\mathcal{N}(r_1) = \{r_2\}$ and $\mathcal{N}(r_K) = \{r_{K-1}\}$. We denote the lower and upper indexed neighbors of rate r by r^- and r^+ given that they exist. Moreover, the set of rates lower than and higher than rate r are denoted by $[r]^-$ and $[r]^+$.

The expected reward function $\mu(p,r)$ exhibits a unimodal structure over the set of rates [5], [6]. For a given transmit power, this structure can be explained via a line graph whose vertices correspond to the rates. More specifically, $\mu(p,r)$ is called *unimodal in the rates* if for any given p there exist a path from any suboptimal rate to the optimal rate along which the expected reward is strictly increasing, i.e., $\forall p \in \mathcal{P}$, $\mu(p,r_1) < \dots < \mu(p,r_p^*)$ and $\mu(p,r_p^*) > \dots > \mu(p,r_K)$. Furthermore, for any given r, $\mu(p,r)$ exhibits a monotone structure over the set of transmit powers, i.e., $\forall r \in \mathcal{R}$, $\mu(p_1,r) \leq \dots \leq \mu(p_P,r)$.

Let $N_{p,r}(t)$ be the number of times rate r was selected by the SU given transmit power p before round t. For a given sequence of transmit powers $p(1),\dots,p(T)$, the (pseudo) regret after the first T rounds is defined as

$$R(T) = \sum_{t=1}^{T} \Big(\mu(p(t), r_{p(t)}^*) - \mu(p(t), r(t))\Big) = \sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{R}} \Delta(p,r) N_{p,r}(T+1). \quad (1)$$

Our goal is to design a learning algorithm for the SU that minimizes the growth rate of the expected regret.

III. The Learning Algorithm

We propose Dynamic Rate Selection via Thompson Sampling (DRS-TS), which is a learning algorithm that takes into account unimodality and monotonicity of $\mu(p,r)$ to minimize the expected regret (pseudocode is given in Algorithm 1). DRS-TS exploits unimodality of $\mu(p,r)$ in rates somewhat similar to UTS [19] and MTS [4]. The main novelty of DRS-TS comes from introduction of transmit power (contextual) information and exploiting the monotonicity of $\mu(p,r)$ in transmit powers. It is important to note that, the learner does not have any control over the transmit power arrivals and the exogenous nature of the arrivals makes efficient learning much more challenging than the non-contextual prior works.

For each power-rate pair (p,r), DRS-TS keeps the counters $N_{p,r}(t)$ and $S_{p,r}(t)$, where $S_{p,r}(t)$ counts the number of successful transmissions in rounds in which transmit power was p and rate r was selected before round t. It also keeps sample mean estimate of the rewards $\hat{\mu}_{p,r}(t)$ that are obtained from rounds in which transmit power was p and rate r was selected prior to the round t.[2]

Algorithm 1 DRS-TS

Input: P, K
Initialize: $t = 1$
Counters: $N_{p,r} = 0, \hat{\mu}_{p,r} = 0, S_{p,r} = 0, b_{p,r} = 0, \forall r \in \mathcal{R}, \forall p \in \mathcal{P}$
while $t \geq 1$ **do**
  Observe transmit power $p(t)$
  $L_{p(t)} = \text{argmax}_{r\in\mathcal{R}}\, \hat{\mu}_{p(t),r}$
  if $\frac{b_{p(t),L_{p(t)}}-1}{3} \in \mathbb{N}$
    $r(t) = L_{p(t)}$
  else
    $\overline{\mathcal{R}} \leftarrow \mathcal{N}(L_{p(t)}) \cup \{L_{p(t)}\}$
    for $r \in \overline{\mathcal{R}}$
      Draw $\phi_{p_j,r}$ from $\pi_{p_j,r}$ in (2), $\forall p_j \in \mathcal{P}$
      $\theta_{p_j,r} = \frac{r}{r_K}\, \phi_{p_j,r}$, $\forall p_j \in \mathcal{P}$
      Find $\overline{\mathcal{M}}_{p(t),r}(t)$ using (3)
      $\underline{\theta}_{p(t),r} = \min_{p'\in\overline{\mathcal{M}}_{p(t),r}(t) \cup \{p(t)\}} \theta_{p',r}$
    end for
    $r(t) = \text{argmax}_{r\in\overline{\mathcal{R}}}\ \underline{\theta}_{p(t),r}$
  end if
  Observe feedback $x_{p(t),r(t)}(t)$ and reward $g_{p(t),r(t)}(t)$
  $b_{p(t),L_{p(t)}} = b_{p(t),L_{p(t)}} + 1$
  $N_{p(t),r(t)} = N_{p(t),r(t)} + 1$
  $\hat{\mu}_{p(t),r(t)} = \frac{\hat{\mu}_{p(t),r(t)}(N_{p(t),r(t)}-1) + g_{p(t),r(t)}(t)}{N_{p(t),r(t)}}$
  $S_{p(t),r(t)} = S_{p(t),r(t)} + x_{p(t),r(t)}(t)$
  $t = t + 1$
end while

The *rate leader* for transmit power $p \in \mathcal{P}$ in round t is defined as the rate with the highest sample mean reward, i.e., $L_p(t) = \text{argmax}_{r\in\mathcal{R}}\, \hat{\mu}_{p,r}(t)$ (ties are broken arbitrarily). Letting $\mathbf{1}(\cdot)$ denote the indicator function, we define

$$b_{p,r}(t) = \sum_{t'=1}^{t-1} \mathbf{1}(p(t') = p, r = L_p(t'))$$

as the number of times rate r was a rate leader when the transmit power was p up to round t.

After observing $p(t)$ in round t, DRS-TS identifies the rate leader $L_{p(t)}(t)$ and calculates $b_{p(t),L_{p(t)}(t)}(t)$. If $(b_{p(t),L_{p(t)}(t)}(t) - 1)/3 \in \mathbb{N}$, then DRS-TS exploits the rate leader to ensure that the current rate leader is selected more often. Similar to [19] and [20], this significantly simplifies the regret analysis. Otherwise, DRS-TS tries to learn the optimal rate for the given transmit power by utilizing unimodality in rates and monotonicity in transmit powers. As a first step, it calculates $\overline{\mathcal{R}}(t) = \mathcal{N}(L_{p(t)}(t)) \cup \{L_{p(t)}(t)\}$. Thanks to unimodality, exploring over $\overline{\mathcal{R}}(t)$ is sufficient for DRS-TS to identify the optimal rate.

[2]When the current round is clear from the context, we suppress the round index.

Selecting from $\overline{\mathcal{R}}(t)$ is performed by using Thompson sampling. The posterior distribution of transmission success probability for all (p,r) in round t is calculated as

$$\pi_{p,r}(t) = Beta\big(1 + S_{p,r}(t),\ 1 + N_{p,r}(t) - S_{p,r}(t)\big) \quad (2)$$

where $Beta(\alpha,\beta)$ is the beta distribution with parameters α and β. A sample drawn from $\pi_{p,r}(t)$ is denoted by $\phi_{p,r}(t)$, and the *throughput sample* is obtained via $\theta_{p,r}(t) = \frac{r}{r_K}\ \phi_{p,r}(t)$. These samples are then used to transfer the knowledge obtained from other transmit powers to the current transmit power. As the first step, we define the monotone neighborhood of $p(t)$ with index $j(t)$, which contains all the contexts greater than the current context as

$$\mathcal{M}_{p(t),r}(t) = \begin{cases} \{p_{j(t)+1}, \ldots, p_P\} & \text{if } j(t) < P \\ \emptyset & \text{otherwise} \end{cases}$$

Monotonicity of $\mu(p,r)$ in transmit powers implies that for a given rate r, it is certain that (p',r) has higher expected reward than (p,r) for all $p' \in \mathcal{M}_{p,r}(t)$. Since the number of observations from a given (p',r) such that $p' \in \mathcal{M}_{p(t),r}(t)$ may be small, to help learning for the current context $\mathcal{M}_{p(t),r}(t)$ is further sorted and only the pairs (p',r) that are observed more than $(p(t),r)$ are selected. Thus, we define the refined monotone neighborhood of $p(t)$ as

$$\overline{\mathcal{M}}_{p(t),r}(t) = \Big\{ p' \in \mathcal{M}_{p(t),r}(t) : N_{p',r}(t) > N_{p(t),r} \Big\}\ . \quad (3)$$

Using the above facts, DRS-TS simply sets its refined throughput sample as

$$\underline{\theta}_{p(t),r}(t) = \min_{p' \in \overline{\mathcal{M}}_{p(t),r}(t)\ \cup\ \{p(t)\}} \theta_{p',r}(t)$$

and then selects the arm with the highest refined sample, i.e., $r(t) = \operatorname{argmax}_{r \in \overline{\mathcal{R}}(t)}\ \underline{\theta}_{p(t),r}(t)$. Note that randomness of the posterior samples ensures sufficient exploration. Finally, at the end of round t, the ACK/NAK feedback $x_{p(t),r(t)}(t)$ is observed, the reward $g_{p(t),r(t)}(t)$ is collected and the sample mean estimates and counters that correspond to the pair $(p(t),r(t))$ are updated.

IV. Regret Analysis of DRS-TS

In this section, we analyze expected value of the regret of DRS-TS given in (1). We build on the non-contextual analysis in [27] to show that DRS-TS achieves logarithmic regret in the contextual setting.

A. Preliminaries

The expected regret can be written as

$$\mathbb{E}[R(T)] = \sum_{p=p_1}^{p_P} \sum_{r \neq r_p^*} \Delta(p,r)\mathbb{E}[N_{p,r}(T+1)]. \quad (4)$$

The standard analysis of frequentist index policies bounds the number of draws of a suboptimal rate by considering two events: i) the optimal rate is under-estimated, and ii) the optimal rate is not under-estimated and the suboptimal rate is selected [5]–[7]. As discussed in [27], we cannot compare throughput sample $\theta_{p,r}(t)$ to the true mean $\mu(p,r)$, due to the fact that $\theta_{p,r}(t)$ is not an optimistic estimate of $\mu(p,r)$. Hence, we compare $\theta_{p,r}(t)$ with a lower confidence bound given as $\mu(p,r) - \beta_{p,r}(t)$, where $\beta_{p,r}(t) := \sqrt{\frac{6\log(b_{p,r_p^*}(t)+1)}{N^*_{p,r}(t)}}$ and $N^*_{p,r}(t)$ is the number of times rate r was selected when the rate leader was r_p^* and the transmit power was p up to round t. Note that $N_{p,r}(t) \geq N^*_{p,r}(t)$.

From KL-UCB-U [6] and Bayes-UCB [28], we have KL-UCB and Bayes-UCB indices at time t as

$$u_{p,r}(t) := \max_{f \in [\hat{\mu}_{p,r}(t), \frac{r}{r_K}]} \{ N_{p,r}(t) d(\frac{\hat{\mu}_{p,r}(t)}{(r/r_K)}, \frac{f}{(r/r_K)}) \leq \log(t) + \log(\log(T)) \}$$

$$q_{p,r}(t) := \frac{r}{r_K}\ Q\Big(1 - \frac{1}{t\log(T)}, \pi_{p,r}(t)\Big)$$

where $d(x,y)$ is the Kullback-Leibler divergence between two Bernoulli distributions x and y, and $Q(a,\pi)$ is the quantile of order a of distribution π, i.e., for $X \sim \pi$, $\mathbb{P}_\pi(X \leq Q(a,\pi)) = a$. It is known that $q_{p,r}(t) \leq u_{p,r}(t)$.

Let $y_{p,r}(t) := \operatorname{argmin}_{\{p' \in \overline{\mathcal{M}}_{p,r}(t)\}} \theta_{p',r}(t)$ denote the target transmit power of power-rate pair (p,r) in round t. Next, we introduce quantities dependent on target transmit power: $M_{p,r}(t)$ is the number of times rate r has been selected when the transmit power was $y_{p,r}(t)$ up to round t, and $\eta_{p,r}(t)$ is the sample mean reward of power-rate pair $(y_{p,r}(t),r)$ up to round t. Similarly, $\vartheta_{p,r}(t)$ is the throughput sample, $o_{p,r}(t)$ is the Bayes-UCB index, and $w_{p,r}(t)$ is the KL-UCB index of power-rate pair $(y_{p,r}(t),r)$ at the beginning of round t.

Let $\tau_p(t)$ denote the round in which transmit power p arrives for the tth time. Let $\tilde{N}_{p,r}(t) := N_{p,r}(\tau_p(t))$, $\tilde{N}^*_{p,r}(t) := N^*_{p,r}(\tau_p(t))$, $\tilde{\mu}_{p,r}(t) := \hat{\mu}_{p,r}(\tau_p(t))$, $\tilde{r}_p(t) := r(\tau_p(t))$, $\tilde{u}_{p,r}(t) := u_{p,r}(\tau_p(t))$, $\tilde{q}_{p,r}(t) := q_{p,r}(\tau_p(t))$, $\tilde{b}_{p,r}(t) := b_{p,r}(\tau_p(t))$, $\tilde{\beta}_{p,r}(t) := \beta_{p,r}(\tau_p(t))$, $\tilde{L}_p(t) := L_p(\tau_p(t))$, $\tilde{\theta}_{p,r}(t) := \theta_{p,r}(\tau_p(t))$, $\tilde{S}_{p,r}(t) := S_{p,r}(\tau_p(t))$, $\tilde{\underline{\theta}}_{p,r}(t) := \underline{\theta}_{p,r}(\tau_p(t))$, $\tilde{y}_{p,r}(t) := y_{p,r}(\tau_p(t))$, $\tilde{M}_{p,r}(t) := M_{p,r}(\tau_p(t))$, $\tilde{\eta}_{p,r}(t) := \eta_{p,r}(\tau_p(t))$, $\tilde{\vartheta}_{p,r}(t) := \vartheta_{p,r}(\tau_p(t))$, $\tilde{o}_{p,r}(t) := o_{p,r}(\tau_p(t))$ and $\tilde{w}_{p,r}(t) := w_{p,r}(\tau_p(t))$. Let $N_p(T)$ represent the number of times the context was p by the end of round T.

Furthermore, we introduce $\tau_{p,r}(s)$ to denote the round in which transmit power is p, rate leader is r_p^* and rate r is selected for the sth time. Let $\tilde{y}^s_{p,r}$ denote the target transmit power in round $\tau_{p,r}(s)$, $\tilde{N}^s_{p,r}$ denote the number of times rate r has been selected in rounds when the transmit power was p and rate leader was r_p^* before round $\tau_{p,r}(s)$, $\tilde{V}_{p,r}(s)$ denote the number of times rate r has been selected when the transmit power was p by the end of round $\tau_{p,r}(s)$, $\tilde{\mu}^s_{p,r}$ denote the sample mean reward calculated from rounds when transmit power was p and rate leader was r_p^* and rate r was selected at the beginning of round $\tau_{p,r}(s)$, and $\tilde{v}^s_{p,r}$ denote the sample mean reward of power-rate pair (p,r) at the beginning of round $\tau_{p,r}(s)$. We have $\tilde{N}^s_{p,r} = (s-1)$, $\tilde{\mu}^s_{p,r} = \frac{1}{(s-1)}\sum_{k=1}^{(s-1)} \tilde{X}^k_{p,r}$ and $\tilde{v}^s_{p,r} = \frac{1}{(\tilde{V}_{p,r}(s)-1)}\sum_{k=1}^{(\tilde{V}_{p,r}(s)-1)} X^k_{p,r}$, where $\tilde{X}^k_{p,r}$ is the reward of rate r when it is selected for kth time when the

transmit power is p and rate leader is r_p^*, and $X_{p,r}^k$ is the reward of rate r when it is selected for kth time when the transmit power is p. For $s = 1$, we use the convention $\tilde{\mu}_{p,r}^s = 0$ and for $\tilde{V}_{p,r}(s) = 1$, $\tilde{v}_{p,r}^s = 0$. Next, we introduce quantities dependent on target transmit power: $\tilde{M}_{p,r}^s$ is the number of times rate r has been selected when the transmit power was $\tilde{y}_{p,r}^s$ before round $\tau_{p,r}(s)$, and $\tilde{\eta}_{p,r}^s$ is the sample mean reward of power-rate pair $(\tilde{y}_{p,r}^s, r)$ at the beginning of round $\tau_{p,r}(s)$. If there is no target transmit power in a certain round, the quantities $\tilde{M}_{p,r}^s$ and $\tilde{\eta}_{p,r}^s$ are zero for $\gamma_{p,r}$ in (8) for that round. We further introduce $\tilde{Z}_{p,r}^{s,*}$ as the number of times rate r_p^* has been selected when transmit power was p and rate leader was r_p^* before round $\tau_{p,r}(s)$, and define $\tilde{g}_{p,r}^{s,*} = \sqrt{6\log(T)/\tilde{Z}_{p,r}^{s,*}}$. Since $\tilde{N}_{p,r_p^*}^*(t) \geq \lfloor \tilde{b}_{p,r_p^*}(t)/3 \rfloor$, the probability that the algorithm has selected the optimal rate for small number of times, when optimal rate is the rate leader is itself small and is given as

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \mathbf{1}\{\tilde{L}_p(t) = r_p^*,\ \tilde{N}_{p,r_p^*}^*(t) \leq (\tilde{b}_{p,r_p^*}(t))^b\}\right] \leq C_b'$$

where $b \in (0,1)$ and $C_b' < \infty$ are constants. This reduces the analysis to rounds when algorithm has seen reasonable number of draws of the optimal rate, and thus the posterior distribution is well concentrated.

We define another term dependent on b as $N_0(b) := \inf\{t \in \mathbb{N} : \log(t)t^b \geq (\sqrt{6}-\sqrt{5})^{-2}\}$. For $\frac{x}{(r/r_K)}, \frac{y}{(r/r_K)} \in [0,1]$, let $d_r(x,y) = d(\frac{x}{(r/r_K)}, \frac{y}{(r/r_K)})$, $d_r^+(x,y) = 0$ if $x \geq y$ and $d_r^+(x,y) = d_r(x,y)$ if $x < y$, and $f(t) := \log(t) + \log(\log(T))$. Let $K_{p,r}^T := \left\lfloor \frac{(1+\epsilon)f(T)}{d_r(\mu(p,r),\mu(p,r_p^*))} \right\rfloor$ for some $\epsilon > 0$. If there exists a rate $r' \in \mathcal{R}$ for which $(r'/r_K) < \mu(p,r_p^*)$, we introduce $\mathcal{N}'(p,r_p^*) := \mathcal{N}(r_p^*)\backslash r'$, otherwise $\mathcal{N}'(p,r_p^*) := \mathcal{N}(r_p^*)$ [5], [8], [18].

B. Main Result

Theorem 1. *For all $\epsilon > 0$, there exists a constant $C > 0$ such that the expected regret of DRS-TS satisfies:*

$$\mathbb{E}[R(T)] \leq \sum_{p=p_1}^{p_P} \sum_{r \in \mathcal{N}'(p,r_p^*)} \left(\gamma_{p,r} \frac{(1+\epsilon)\Delta(p,r)}{d\left(\frac{\mu(p,r)}{r/r_K}, \frac{\mu(p,r_p^*)}{r/r_K}\right)}\right)\log(T) + O(\log(\log(T))) + C$$

where $\gamma_{p,r} = \frac{\sum_{s=1}^{K_{p,r}^T+1} \mathbb{P}\left(\tilde{M}_{p,r}^s d_r^+(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - \tilde{g}_{p,r}^{s,*}) < f(T)\right)}{K_{p,r}^T+1} \in [0,1]$.

The term $\gamma_{p,r}$ depends on how well power-rate pair (p,r) has exploited its monotone neighborhood. For negligible selections of rate r by the target transmit power, $\gamma_{p,r}$ is high and close to 1. As the target transmit power increases selections of rate r, $\gamma_{p,r}$ decreases, and the regret of DRS-TS decreases.

C. Proof of Theorem 1

For p such that $N_p(T) > 0$ and $r \neq r_p^*$, the expectation in (4) is decomposed into two terms as in [19, Theorem 2]: $\mathbb{E}[N_{p,r}(T+1)] = \mathbb{E}[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{r}_p(t) = r\}] = \mathbb{E}[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) \neq r_p^*, \tilde{r}_p(t) = r\} + \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r\}]$. We say that a rate r is suboptimal rate for a given transmit power p if $\Delta(p,r) > 0$. Since only the rate leader and its neighbors are explored, when the rate leader is r_p^* we only select from rates that lie in $\mathcal{N}(r_p^*) \cup \{r_p^*\}$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}[R(T)] \\
&= \underbrace{\sum_{p=p_1}^{p_P} \left(\sum_{r \neq r_p^*} \Delta(p,r) \mathbb{E}\left[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) \neq r_p^*, \tilde{r}_p(t) = r\}\right]\right.}_{(\mathbf{A})} \\
&\left. + \sum_{r \in \mathcal{N}(r_p^*)} \Delta(p,r) \mathbb{E}\left[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r\}\right]\right). \quad (5)
\end{aligned}$$

Note that $(\mathbf{A}) \leq \sum_{p=p_1}^{p_P} \sum_{r \neq r_p^*} \mathbb{E}[b_{p,r}(T+1)]$. Similar to [21, Theorem 5], a suboptimal rate r can be the rate leader for a given transmit power p only for a small number of times, and thus, we have $\mathbb{E}[b_{p,r}(T+1)] \leq C_1$, where C_1 is a positive constant.

For $\mathcal{N}(r_p^*) \neq \mathcal{N}'(p,r_p^*)$, let $r' = \mathcal{N}(r_p^*)\backslash\mathcal{N}'(p,r_p^*)$. We decompose second term in (5) as

$$\begin{aligned}
&= \underbrace{\Delta(p,r') \mathbb{E}\left[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r'\}\right]}_{(\mathbf{B})} \\
&+ \sum_{r \in \mathcal{N}'(p,r_p^*)} \Delta(p,r) \underbrace{\mathbb{E}\left[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r\}\right]}_{(\mathbf{C})}.
\end{aligned}$$

Note that $(\mathbf{B})$ is neglected when $\mathcal{N}(r_p^*)\backslash\mathcal{N}'(p,r_p^*) = \emptyset$. We have

$$\begin{aligned}
(\mathbf{B}) &\leq \mathbb{E}\Big[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \frac{r'}{r_K} \geq \underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big] \\
&\quad + \mathbb{E}\Big[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r', \frac{r'}{r_K} < \underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big] \\
&\leq \mathbb{E}\Big[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \frac{r'}{r_K} \geq \underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big] \leq PC_a.
\end{aligned}$$

The above holds since $\underline{\tilde{\theta}}_{p,r'}(t) \in [0, \frac{r'}{r_K}]$, and for event $\{\tilde{L}_p(t) = r_p^*, \tilde{r}_p(t) = r', \frac{r'}{r_K} < \underline{\tilde{\theta}}_{p,r_p^*}(t)\}$ to happen we need $\underline{\tilde{\theta}}_{p,r'}(t) \geq \underline{\tilde{\theta}}_{p,r_p^*}(t)$ which cannot happen when $\underline{\tilde{\theta}}_{p,r_p^*}(t) > \frac{r'}{r_K}$. C_a is independent of T and comes from [29, Lemma 9], which bounds underestimation of Thompson sample for an optimal arm (rate) from a fixed distance, i.e., $\mu(p,r_p^*) - \delta = r'/r_K$.

For $r \in \mathcal{N}'(p,r_p^*)$, we have $(\mathbf{C}) \leq (\mathbf{D}) + (\mathbf{E})$, where

$$\begin{aligned}
(\mathbf{D}) := \mathbb{E}\Big[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*, \\
\mu(p,r_p^*) - \tilde{\beta}_{p,r_p^*}(t) > \underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big]
\end{aligned}$$

$$(\mathbf{E}) := \mathbb{E}\Big[\sum_{t=1}^{N_p(T)} \mathbf{1}\{\tilde{L}_p(t) = r_p^*,\ \tilde{r}_p(t) = r, \mu(p, r_p^*) - \tilde{\beta}_{p,r_p^*}(t) \le \underline{\tilde{\theta}}_{p,r_p^*}(t)\}\Big].$$

Next, we bound (**E**). Let $Q_t := \{\tilde{\theta}_{p,r}(t) \le \tilde{q}_{p,r}(t), \tilde{\vartheta}_{p,r}(t) \le \tilde{o}_{p,r}(t)\}$. Note that $\tilde{L}_p(t) = r_p^*$ and $\tilde{r}_p(t) = r$ together imply that $\underline{\tilde{\theta}}_{p,r_p^*}(t) \le \underline{\tilde{\theta}}_{p,r}(t)$ and recall that $\underline{\tilde{\theta}}_{p,r}(t) = \min\{\tilde{\theta}_{p,r}(t),\ \tilde{\vartheta}_{p,r}(t)\}$. Thus, we have (**E**) $\le$

$$\begin{aligned}
&\sum_{t=1}^{N_p(T)} \Big(\mathbb{P}(\tilde{L}_p(t)=r_p^*, \tilde{r}_p(t)=r, \mu(p,r_p^*) - \tilde{\beta}_{p,r_p^*}(t) \le \tilde{\theta}_{p,r}(t), \\
&\qquad \mu(p,r_p^*) - \tilde{\beta}_{p,r_p^*}(t) \le \tilde{\vartheta}_{p,r}(t), Q_t) + \mathbb{P}(Q_t^c)\Big) \\
&\le \sum_{t=1}^{N_p(T)} \mathbb{P}(\tilde{L}_p(t)=r_p^*, \tilde{r}_p(t)=r, \mu(p,r_p^*) - \tilde{\beta}_{p,r_p^*}(t) \le \tilde{u}_{p,r}(t), \\
&\qquad \mu(p,r_p^*) - \tilde{\beta}_{p,r_p^*}(t) \le \tilde{w}_{p,r}(t)) + \sum_{t=1}^{N_p(T)} \mathbb{P}(Q_t^c). \qquad (6)
\end{aligned}$$

The last inequality comes from the fact that the event Q_t and fact $q_{p,r}(t) \le u_{p,r}(t)$ together ensure that $\tilde{\theta}_{p,r}(t) \le \tilde{q}_{p,r}(t) \le \tilde{u}_{p,r}(t)$ and $\tilde{\vartheta}_{p,r}(t) \le \tilde{o}_{p,r}(t) \le \tilde{w}_{p,r}(t)$. Since, the samples $\tilde{\theta}_{p,r}(t)$ and $\tilde{\vartheta}_{p,r}(t)$ are not very likely to exceed the corresponding quantiles of the posterior distribution [27], we have

$$\begin{aligned}
\sum_{t=1}^{N_p(T)} \mathbb{P}(Q_t^c) &= \sum_{t=1}^{N_p(T)} \mathbb{P}(\tilde{\theta}_{p,r}(t) > \tilde{q}_{p,r}(t) \cup \tilde{\vartheta}_{p,r}(t) > \tilde{o}_{p,r}(t)) \\
&\le P \sum_{t=1}^{T} \frac{1}{t\log(T)} \le 2P.
\end{aligned}$$

Let $\tau_{p,t'}^*$ represent the round at which rate r_p^* is the rate leader for t'th time, when the transmit power is p. We bound the first term in (6) for $r \in \mathcal{N}'(p, r_p^*)$ as

$$\begin{aligned}
&\le \sum_{t=1}^{N_p(T)} \mathbb{P}(\tilde{L}_p(t) = r_p^*,\ \tilde{r}_p(t) = r,\ \tilde{N}_{p,r_p^*}^*(t) > (\tilde{b}_{p,r_p^*}(t))^b, \\
&\quad \mu(p,r_p^*) - \tilde{\beta}_{p,r_p^*}(t) \le \tilde{u}_{p,r}(t), \mu(p,r_p^*) - \tilde{\beta}_{p,r_p^*}(t) \le \tilde{w}_{p,r}(t)) \\
&\quad + \sum_{t=1}^{N_p(T)} \mathbb{P}\big(\tilde{L}_p(t) = r_p^*, \tilde{N}_{p,r_p^*}^*(t) \le (\tilde{b}_{p,r_p^*}(t))^b\big) \\
&\le \sum_{t'=1}^{N_p(T)} \sum_{t=1}^{N_p(T)} \mathbb{P}(\tilde{L}_p(t) = r_p^*,\ \tilde{b}_{p,r_p^*}(t) = t'-1, \\
&\quad \tilde{r}_p(t) = r, \tilde{N}_{p,r_p^*}^*(t) > (t'-1)^b, \\
&\quad \mu(p,r_p^*) - \sqrt{\frac{6\log(t')}{\tilde{N}_{p,r_p^*}^*(t)}} \le \tilde{u}_{p,r}(t), \\
&\quad \mu(p,r_p^*) - \sqrt{\frac{6\log(t')}{\tilde{N}_{p,r_p^*}^*(t)}} \le \tilde{w}_{p,r}(t)) + C_b' \\
&\le \sum_{t'=1}^{N_p(T)} \mathbb{P}(r(\tau_{p,t'}^*) = r, N_{p,r_p^*}^*(\tau_{p,t'}^*) > (t'-1)^b, \\
&\quad \mu(p,r_p^*) - \sqrt{\frac{6\log(t')}{N_{p,r_p^*}^*(\tau_{p,t'}^*)}} \le u_{p,r}(\tau_{p,t'}^*), \\
&\quad \mu(p,r_p^*) - \sqrt{\frac{6\log(t')}{N_{p,r_p^*}^*(\tau_{p,t'}^*)}} \le w_{p,r}(\tau_{p,t'}^*)) + C_b'. \qquad (7)
\end{aligned}$$

Let $\alpha_{p,r_p^*}^{t'} = \sqrt{6\log(t')/N_{p,r_p^*}^*(\tau_{p,t'}^*)}$. Since $u_{p,r}(\tau_{p,t'}^*) \ge \mu(p,r_p^*) - \alpha_{p,r_p^*}^{t'}$, we have

$$\begin{aligned}
&N_{p,r}^*(\tau_{p,t'}^*) d_r^+(\hat{\mu}_{p,r}(\tau_{p,t'}^*), \mu(p,r_p^*) - \alpha_{p,r_p^*}^{t'}) \\
&\qquad \le N_{p,r}(\tau_{p,t'}^*) d_r(\hat{\mu}_{p,r}(\tau_{p,t'}^*), u_{p,r}(\tau_{p,t'}^*)) \\
&\qquad \le \log(T) + \log(\log(T)) = f(T).
\end{aligned}$$

Similarly, condition $w_{p,r}(\tau_{p,t'}^*) \ge \mu(p,r_p^*) - \alpha_{p,r_p^*}^{t'}$ implies that

$$\begin{aligned}
&M_{p,r}(\tau_{p,t'}^*) d_r^+(\eta_{p,r}(\tau_{p,t'}^*), \mu(p,r_p^*) - \alpha_{p,r_p^*}^{t'}) \\
&\qquad \le M_{p,r}(\tau_{p,t'}^*) d_r(\eta_{p,r}(\tau_{p,t'}^*), w_{p,r}(\tau_{p,t'}^*)) \le f(T).
\end{aligned}$$

Using the inequalities above, we bound the summation term in (7) as

$$\begin{aligned}
&\sum_{t'=1}^{N_p(T)} \mathbb{P}\big(r(\tau_{p,t'}^*) = r, N_{p,r_p^*}^*(\tau_{p,t'}^*) > (t'-1)^b, \\
&\quad \mu(p,r_p^*) - \alpha_{p,r_p^*}^{t'} \le u_{p,r}(\tau_{p,t'}^*), \mu(p,r_p^*) - \alpha_{p,r_p^*}^{t'} \le w_{p,r}(\tau_{p,t'}^*)\big) \\
&\le E\Bigg[\sum_{t'=1}^{N_p(T)} \mathbf{1}\Big(r(\tau_{p,t'}^*) = r, N_{p,r_p^*}^*(\tau_{p,t'}^*) > (t'-1)^b, \\
&\quad N_{p,r}^*(\tau_{p,t'}^*) d_r^+(\hat{\mu}_{p,r}(\tau_{p,t'}^*), \mu(p,r_p^*) - \alpha_{p,r_p^*}^{t'}) \le f(T), \\
&\quad M_{p,r}(\tau_{p,t'}^*) d_r^+(\eta_{p,r}(\tau_{p,t'}^*), \mu(p,r_p^*) - \alpha_{p,r_p^*}^{t'}) \le f(T)\Big)\Bigg].
\end{aligned}$$

We introduce s and bound the sum above as follows with the fact that $N_p(T) \le T$:

$$\begin{aligned}
&\le E\Bigg[\sum_{t'=1}^{T} \sum_{s=1}^{t'} \mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*) = (s-1), \\
&\quad r(\tau_{p,t'}^*) = r, \tilde{Z}_{p,r}^{s,*} > (t'-1)^b, \\
&\quad (s-1) d_r^+\big(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - \sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big) \le f(T), \\
&\quad \tilde{M}_{p,r}^s d_r^+\big(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - \sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big) \le f(T)\big)\Bigg].
\end{aligned}$$

By change of variables, and separating outer sum into two intervals, we have

$$\begin{aligned}
&\le E\Bigg[\sum_{s=1}^{K_{p,r}^T+1} \sum_{t'=s}^{T} \mathbf{1}\big(N_{p,r}^*(\tau_{p,t'}^*) = (s-1), \\
&\quad r(\tau_{p,t'}^*) = r, \tilde{Z}_{p,r}^{s,*} > (t'-1)^b, \\
&\quad (s-1) d_r^+\big(\tilde{v}_{p,r}^s, \mu(p,r_p^*) - \sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big) \le f(T), \\
&\quad \tilde{M}_{p,r}^s d_r^+\big(\tilde{\eta}_{p,r}^s, \mu(p,r_p^*) - \sqrt{6\log(t')/\tilde{Z}_{p,r}^{s,*}}\ \big) \le f(T)\big)
\end{aligned}$$

$$+\sum_{s=K^T_{p,r}+2}^{T}\sum_{t'=s}^{T}\mathbf{1}\big(N^*_{p,r}(\tau^*_{p,t'})=(s-1),$$
$$r(\tau^*_{p,t'})=r,\tilde{Z}^{s,*}_{p,r}>(t'-1)^b,$$
$$(s-1)d_r^+\big(\tilde{v}^s_{p,r},\mu(p,r^*_p)-\sqrt{6\log(t')/\tilde{Z}^{s,*}_{p,r}}\,\big)\le f(T),$$
$$\tilde{M}^s_{p,r}d_r^+\big(\tilde{\eta}^s_{p,r},\mu(p,r^*_p)-\sqrt{6\log(t')/\tilde{Z}^{s,*}_{p,r}}\,\big)\le f(T)\big)\Bigg].$$

For first summation we use $\sqrt{6\log(t')/\tilde{Z}^{s,*}_{p,r}}\le\tilde{g}^{s,*}_{p,r}$ because $t'\le T$ and $\tilde{g}^{s,*}_{p,r}=\sqrt{6\log(T)/\tilde{Z}^{s,*}_{p,r}}$. For second summation we utilize the condition $\tilde{Z}^{s,*}_{p,r}>(t'-1)^b$ and have $\sqrt{6\log(t')/\tilde{Z}^{s,*}_{p,r}}<\sqrt{6\log(t')/(t'-1)^b}$. We introduce $h_t=\sqrt{\frac{6\log(t+1)}{(t)^b}}$, and thus

$$\le E\Bigg[\sum_{s=1}^{K^T_{p,r}+1}\sum_{t'=s}^{T}\mathbf{1}\big(N^*_{p,r}(\tau^*_{p,t'})=(s-1),r(\tau^*_{p,t'})=r,$$
$$(s-1)d_r^+(\tilde{v}^s_{p,r},\mu(p,r^*_p)-\tilde{g}^{s,*}_{p,r})\le f(T),$$
$$\tilde{M}^s_{p,r}d_r^+(\tilde{\eta}^s_{p,r},\mu(p,r^*_p)-\tilde{g}^{s,*}_{p,r})\le f(T)\big)$$
$$+\sum_{s=K^T_{p,r}+2}^{T}\sum_{t'=s}^{T}\mathbf{1}\big(N^*_{p,r}(\tau^*_{p,t'})=(s-1),r(\tau^*_{p,t'})=r,$$
$$(s-1)d_r^+(\tilde{v}^s_{p,r},\mu(p,r^*_p)-h_{t'-1})\le f(T),$$
$$\tilde{M}^s_{p,r}d_r^+(\tilde{\eta}^s_{p,r},\mu(p,r^*_p)-h_{t'-1})\le f(T)\big)\Bigg].$$

Note that $q\mapsto d_r^+(p,q)$ is nondecreasing, and for large enough t' (i.e., $t'>e^{1/b}$), $t'\mapsto h_{t'}$ is decreasing. Thus, for the second term above, for T such that $K^T_{p,r}>e^{1/b}$, we have $h_{t'-1}\le h_{K^T_{p,r}+1}<h_{K^T_{p,r}}$ due to the fact that $t'-1\ge s-1\ge K^T_{p,r}+1$. We separate t' and s terms and obtain

$$\le E\Bigg[\sum_{s=1}^{K^T_{p,r}+1}\mathbf{1}\big((s-1)d_r^+(\tilde{v}^s_{p,r},\mu(p,r^*_p)-\tilde{g}^{s,*}_{p,r})\le f(T),$$
$$\tilde{M}^s_{p,r}d_r^+(\tilde{\eta}^s_{p,r},\mu(p,r^*_p)-\tilde{g}^{s,*}_{p,r})\le f(T)\big)$$
$$\sum_{t'=s}^{T}\mathbf{1}\big(N^*_{p,r}(\tau^*_{p,t'})=(s-1),r(\tau^*_{p,t'})=r\big)$$
$$+\sum_{s=K^T_{p,r}+2}^{T}\mathbf{1}\big((s-1)d_r^+(\tilde{v}^s_{p,r},\mu(p,r^*_p)-h_{K^T_{p,r}})\le f(T),$$
$$\tilde{M}^s_{p,r}d_r^+(\tilde{\eta}^s_{p,r},\mu(p,r^*_p)-h_{K^T_{p,r}})\le f(T)\big)$$
$$\sum_{t'=s}^{T}\mathbf{1}\big(N^*_{p,r}(\tau^*_{p,t'})=(s-1),r(\tau^*_{p,t'})=r\big)\Bigg].$$

Since, $\sum_{t'=s}^{T}\mathbf{1}\big(N^*_{p,r}(\tau^*_{p,t'})=(s-1),r(\tau^*_{p,t'})=r\big)\le 1$ for every $s\in[1,T]$, and using the fact $\mathbb{P}(A,B)\le\mathbb{P}(A)$ for events A and B, we have

$$\le\sum_{s=1}^{K^T_{p,r}+1}\mathbb{P}\big(\tilde{M}^s_{p,r}d_r^+(\tilde{\eta}^s_{p,r},\mu(p,r^*_p)-\tilde{g}^{s,*}_{p,r})\le f(T)\big)$$
$$+\sum_{s=K^T_{p,r}+2}^{T}\mathbb{P}\big((s-1)d_r^+(\tilde{v}^s_{p,r},\mu(p,r^*_p)-h_{K^T_{p,r}})\le f(T)\big).\quad(8)$$

We let $\gamma_{p,r}:=\frac{\sum_{s=1}^{K^T_{p,r}+1}\mathbb{P}\left(\tilde{M}^s_{p,r}d_r^+(\tilde{\eta}^s_{p,r},\mu(p,r^*_p)-\tilde{g}^{s,*}_{p,r})<f(T)\right)}{K^T_{p,r}+1}$, and write the first term in (8) as $\gamma_{p,r}(K^T_{p,r}+1)$.

We bound the second term in (8) as

$$\sum_{s=K^T_{p,r}+2}^{T}\mathbb{P}((s-1)d_r^+(\tilde{v}^s_{p,r},\mu(p,r^*_p)-h_{K^T_{p,r}})\le f(T))$$
$$\le\sum_{s=K^T_{p,r}+2}^{T}\mathbb{P}((K^T_{p,r}+1)d_r^+(\tilde{v}^s_{p,r},\mu(p,r^*_p)-h_{K^T_{p,r}})\le f(T))$$
$$\le\sum_{s=K^T_{p,r}+2}^{T}\mathbb{P}\Big(d_r^+(\tilde{v}^s_{p,r},\mu(p,r^*_p)-h_{K^T_{p,r}})$$
$$\le\frac{d_r(\mu(p,r),\mu(p,r^*_p))}{1+\epsilon}\Big)\le C_2$$

where C_2 is a constant from [27, Lemma 2] and utilizing the fact that $\mathbb{P}(\tilde{V}_{p,r}(s)\ge s)=1$.

Finally, we bound (**D**):

$$(\mathbf{D})\le\sum_{t'=1}^{N_p(T)}\mathbb{P}(\mu(p,r^*_p)-\sqrt{\frac{6\log(t')}{N^*_{p,r^*_p}(\tau^*_{p,t'})}}>\theta_{p,r^*_p}(\tau^*_{p,t'}))$$
$$+\sum_{t'=1}^{N_p(T)}\mathbb{P}(\mu(p,r^*_p)-\sqrt{\frac{6\log(t')}{N^*_{p,r^*_p}(\tau^*_{p,t'})}}>\vartheta_{p,r^*_p}(\tau^*_{p,t'}))$$
$$\le P(N_0(b)+3+C'_b)=\ C_3$$

where the last inequality comes from [27, Lemma 1]. Using above bounds, we obtain

$$\mathbb{E}[R(T)]=\sum_{p=p_1}^{p_P}\sum_{r\neq r^*_p}\Delta(p,r)\mathbb{E}[N_{p,r}(T+1)]$$
$$\le\sum_{p=p_1}^{p_P}\sum_{r\in\mathcal{N}'(p,r^*_p)}\Big(\gamma_{p,r}\frac{\Delta(p,r)(1+\epsilon)}{d\left(\frac{\mu(p,r)}{(r/r_K)},\frac{\mu(p,r^*_p)}{(r/r_K)}\right)}\Big)\log(T)$$
$$+O(\log(\log(T)))+C$$

where $C<\infty$ is a constant given as $C:=(P)(R-1)(C_1)+P^2C_a+\sum_{p=p_1}^{p_P}\sum_{r\in\mathcal{N}'(p,r^*_p)}\Delta(p,r)\Big(C_2+C_3+\gamma_{p,r}+2P+C'_b\Big)$.

V. Illustrative Results

In this section, we numerically evaluate the performance of DRS-TS for dynamic rate selection under time-varying transmit power and compare its performance with the state-of-the-art algorithms.

A. Competitor Learning Algorithms

1) *Dynamic Rate Selection via Thompson Sampling Without Contexts* (DRS-TS-NC): This is the non-contextual version of DRS-TS. It decouples the rate from throughput and exploits unimodality of the expected reward similar to MTS [4] and UTS [19].

TABLE I
COMPARISON OF DRS-TS WITH STATE-OF-THE-ART ALGORITHMS

Algorithm	Arm unimodality	Contexts	Context monotonicity	The Expected Regret
DRS-TS-NC	Yes	No	No	$O(\lvert\mathcal{N}'(r^*)\rvert \log(T))$
CUCB [30]	No	Yes	No	$O(\sum_{p\in\mathcal{P}} \sum_{r\neq r_p^*} \log(T))$
DRS-KLUCB [18]	Yes	Yes	Yes	$O(\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \gamma_{p,r} \log(T))$, $\gamma_{p,r} \in [0,1]$
CUTS [19]	Yes	Yes	No	$O(\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}(r_p^*)} \log(T))$
DRS-TS-NU	Yes	Yes	No	$O(\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \log(T))$
DRS-TS (our work)	Yes	Yes	Yes	$O(\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \gamma_{p,r} \log(T))$, $\gamma_{p,r} \in [0,1]$

2) *Contextual Upper Confidence Bound* (CUCB): This runs a separate instance of UCB1 [30] for each transmit power. It does not use rate unimodality.

3) *Dynamic Rate Selection via Kullback–Leibler Upper Confidence Bound* (DRS-KLUCB): This is the frequentist variant of DRS-TS, which is a simplified version of CUL [18], where the modified neighborhood is set beforehand similar to DRS-TS. This benchmark is used to test if the Bayesian approach is superior to the frequentist approach in practice.

4) *Contextual Unimodal Thompson Sampling* (CUTS): This is the contextual version of unimodal Thompson sampling (UTS) proposed in [19]. It runs a separate instance of UTS for each transmit power.

5) *Dynamic Rate Selection via Thompson Sampling without contextual unimodality* (DRS-TS-NU): This runs a separate instance of DRS-TS-NC for each transmit power. It is a variant of DRS-TS that decouples the rate from the throughput, exploits the unimodality over rates and contextual information but ignores the monotonicity over the contexts. This benchmark evaluates the effect of exploiting the monotonicity over transmit powers on the regret.

B. Regret and Complexity Comparison

The expected regret of DRS-TS is compared with the other learning algorithms in Table I. It is evident that exploiting unimodality reduces the search space to the neighborhood of the optimal rate. Additionally, exploiting the context monotonicity introduces the context dependent factor $\gamma_{p,r} \in [0,1]$, which ensures the fact that $\sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \gamma_{p,r} \log(T) \leq \sum_{p\in\mathcal{P}} \sum_{r\in\mathcal{N}'(p,r_p^*)} \log(T)$. While the frequentist analogue of DRS-TS achieves a regret bound similar to that of DRS-TS, our experimental results illustrate that the performance of DRS-TS is superior in practice.

DRS-TS-NC has the lowest complexity, since it neglects the context. However, it performs poorly. Contextual algorithms have higher complexity but they achieve a better performance than the non-contextual algorithms. DRS-TS has the highest computational complexity, since it needs to refine the throughput sample via exploiting the neighborhood contexts, whereas the performance achieved is greater than all of the competitor algorithms. Nevertheless, computational complexity is linear in terms of number of rates and transmit powers, and hence, in practice the algorithm can be deployed in a device with limited computational resources.

TABLE II
THROUGHPUT OF POWER-RATE PAIRS.

Throughput	r_1	r_2	r_3	r_4
p_1	**0.1233**	0.0602	0.0042	0.0000
p_2	**0.1236**	0.0623	0.0052	0.0000
p_3	**0.1247**	0.0630	0.0052	0.0000
p_4	**0.1264**	0.0658	0.0052	0.0000
p_5	**0.1288**	0.0672	0.0052	0.0000
p_6	**0.1292**	0.0699	0.0073	0.0000
p_7	0.1970	**0.2625**	0.1558	0.0305
p_8	0.1977	**0.2652**	0.1589	0.0332
p_9	0.1988	**0.2673**	0.1610	0.0332
p_{10}	0.1994	**0.2722**	0.1662	0.0332
p_{11}	0.2005	**0.2742**	0.1662	0.0332
p_{12}	0.2008	**0.2742**	0.1735	0.0346
p_{13}	0.2396	0.4301	**0.5287**	0.4571
p_{14}	0.2396	0.4307	**0.5287**	0.4584
p_{15}	0.2400	0.4321	**0.5287**	0.4584
p_{16}	0.2403	0.4328	**0.5308**	0.4626
p_{17}	0.2403	0.4335	**0.5319**	0.4640
p_{18}	0.2403	0.4356	**0.5319**	0.4709

C. Experimental Setup

We consider a single-link system with 4 available rates and 18 transmit powers. The rate set is $\{2, 4, 6, 8\}$ bits per symbol (bps), which corresponds to modulation schemes QPSK, 16QAM, 64QAM, and 256QAM. Contexts (transmit powers) correspond to average signal-to-noise-ratio (SNR) values ranging in $[2, 25]$dBs, and we set $T = 7.2\times10^4$. A snapshot of throughput calculated for various power-rate pairs is provided in Table II. Packet success probabilities are obtained via sending packets over a *tapped delay line* (TDL) channel with Rayleigh Fading using 5G toolbox in MATLAB. For each experiment run, Bernoulli random numbers are generated by using the obtained probabilities and random rewards are generated by multiplying the obtained random number with the corresponding rate. Results are averaged over 50 runs to reduce the effect of randomness due to rate selections, transmit power arrivals and reward generation. We evaluate performance of the algorithms under three different types of context arrivals that are explained in the following subsections.

D. Type I Arrivals

We consider a sequence of ordered contexts $\{p_{18}, \ldots, p_1\}$, where each context arrives for a block of $T/18$ samples. Therefore, the expected reward for rate $r \in \mathcal{R}$ decreases monotonically, i.e., $\mu(p(t+1), r) \leq \mu(p(t), r)$. The monotonicity of throughput over transmit power is maximally utilized for Type I arrivals due to the fact that contexts with higher values of

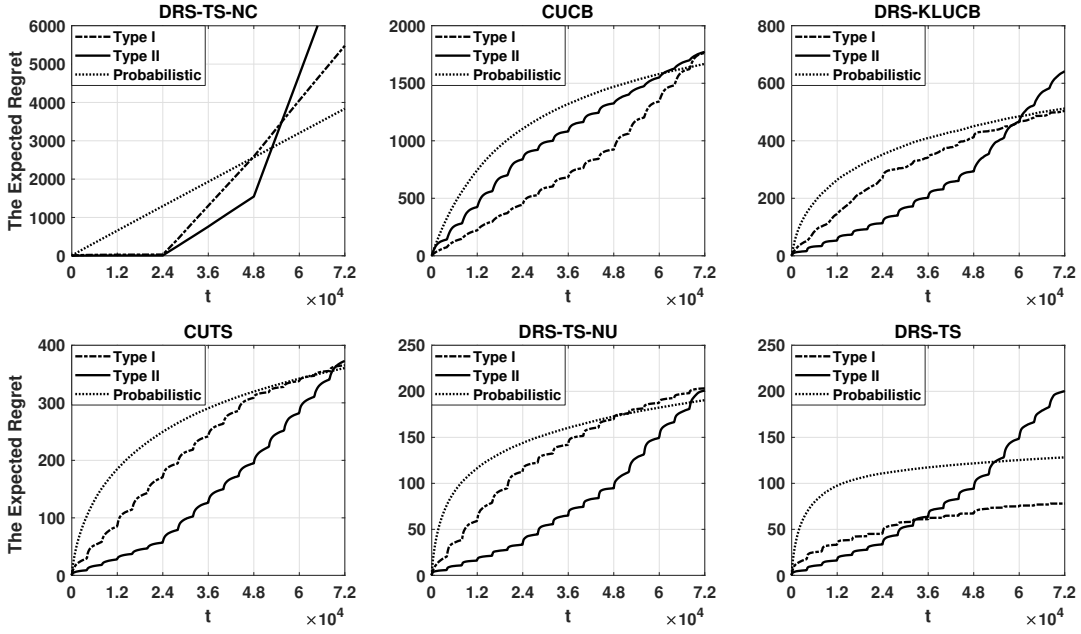

Fig. 1. The expected regrets of DRS-TS-NC, CUCB, DRS-KLUCB, CUTS, DRS-TS-NU and DRS-TS.

expected reward arrive early, and monotone neighborhood for most of the contexts provides a refined *throughput sample*. Fig. 1 compares the performance of DRS-TS-NC, CUCB, DRS-KLUCB, CUTS, DRS-TS-NU and DRS-TS for Type I arrivals. DRS-TS-NC tries to find a global optimal rate for all transmit powers, and hence, it incurs the largest regret. Although CUCB obtains the optimal rate for each transmit power, it still incurs a large regret due to not exploiting unimodality in the rates. Another point worth noting is that CUTS and DRS-TS-NU outperforms DRS-KLUCB, which provides evidence for the fact that the Bayesian approach results in faster learning than the frequentist approach. Also DRS-TS-NU outperforms CUTS, since DRS-TS-NU decouples the rate and packet success probability in addition to rate unimodality exploitation. Finally, DRS-TS achieves considerably smaller regret than DRS-TS-NC since it additionally utilizes the transmit power monotonicity. Also DRS-TS outperforms its frequentist counterpart DRS-KLUCB with a significant margin.

E. Type II Arrivals

We again consider a sequence of ordered contexts $\{p_1, \ldots, p_{18}\}$, where each context arrives for a block of $T/18$ samples. However, in this scenario the expected reward for rate $r \in \mathcal{R}$ increases monotonically i.e., $\mu(p(t+1), r) \geq \mu(p(t), r)$. The monotonicity of throughput over transmit power is not utilized for Type II arrivals, due to the fact that contexts with lower expected rewards arrive first, and hence, the monotone neighborhood for all of the contexts is an empty set. In Fig. 1 we show that similar to Type I arrivals, DRS-TS-NU outperforms DRS-TS-NC, CUCB, DRS-KLUCB and CUTS thanks to exploiting the rate decomposition, contextual information and rate unimodality. However, DRS-TS behaves similarly to DRS-TS-NU, since transmit power monotonicity does not have impact for Type II arrivals. Similar to Type I arrivals, a staircase pattern is observed for the contextual algorithms in Fig. 1. This pattern is due to the fact that the context arrivals follow a monotone trend. In contrast, the regret due to exploring suboptimal arms is growing logarithmic in time as seen by the form of each staircase step.

F. Probabilistic Arrivals

For this case, there are three sets of 6 transmit powers each for which the optimal rate is same as given in Table II. Each time a set is selected uniformly at random and then sorted (descending) transmit powers inside the selected set are selected with the following probabilities $\{\frac{6}{21}, \frac{5}{21}, \frac{4}{21}, \frac{3}{21}, \frac{2}{21}, \frac{1}{21}\}$. The monotonicity of throughput over transmit powers is considerably utilized for probabilistic arrivals, due to the fact that contexts with higher values of expected reward arrive more frequently within each set, and monotone neighborhood for most of the contexts can provide a refined *throughput sample*. The utilization of transmit power monotonicity in probabilistic case is in between the two extreme cases (Type I and II arrivals). Fig. 1 shows that similar to Type I and Type II arrivals, DRS-TS-NU outperforms DRS-TS-NC, CUCB, DRS-KLUCB and CUTS. DRS-TS outperforms DRS-TS-NU since transmit power monotonicity is considerably utilized.

VI. Conclusion

In this paper, we considered the problem of rate selection under time-varying transmit power over an mmWave channel. We proposed a Bayesian learning algorithm, called DRS-TS, that exploits the structure of the throughput in rates as well as in transmit powers efficiently. We proved upper bounds on the regret of DRS-TS and presented experiments that compare the performance of DRS-TS with the state-of-the-art. Our results indicate that DRS-TS results in significant performance gains across a variety of context arrival patterns.

REFERENCES

[1] Federal Communications Commission, "FCC adopts rules to facilitate next generation wireless technologies," *FCC, July 14*, 2016.

[2] K. Haneda, L. Tian, H. Asplund, J. Li, Y. Wang, D. Steer, C. Li, T. Balercia, S. Lee, Y. Kim, *et al.*, "Indoor 5G 3GPP-like channel models for office and shopping mall environments," in *Proc. IEEE Int. Conf. Commun. Workshops*, pp. 694–699, May 2016.

[3] K. Haneda, J. Zhang, L. Tan, G. Liu, Y. Zheng, H. Asplund, J. Li, Y. Wang, D. Steer, C. Li, *et al.*, "5G 3GPP-like channel models for outdoor urban microcellular and macrocellular environments," in *Proc. 83rd IEEE Veh. Technol. Conf.*, pp. 1–7, May 2016.

[4] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-complexity, low-regret link rate selection in rapidly-varying wireless channels," in *Proc. IEEE Conf. Comput. Commun.*, pp. 540–548, Apr. 2018.

[5] R. Combes, A. Proutiere, D. Yun, J. Ok, and Y. Yi, "Optimal rate sampling in 802.11 systems," in *Proc. IEEE Conf. Comput. Commun.*, pp. 2760–2767, Apr. 2014.

[6] 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.

[7] R. Combes, J. Ok, A. Proutiere, D. Yun, and Y. Yi, "Optimal rate sampling in 802.11 systems: Theory, design, and implementation," *IEEE Trans. Mobile Comput.*, vol. 18, no. 5, pp. 1145–1158, 2018.

[8] H. Gupta, A. Eryilmaz, and R. Srikant, "Link rate selection using constrained Thompson sampling," in *Proc. IEEE Conf. Comput. Commun.*, pp. 739–747, 2019.

[9] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *Proc. IEEE Conf. Comput. Commun.*, pp. 2393–2401, Apr. 2018.

[10] D. E. Papanikolaou, N. E. Papanikolaou, G. T. Pitsiladis, A. D. Panagopoulos, and Ph. Constantinou, "Spectrum sensing in mm-Wave cognitive radio networks under rain fading," in *Proc. 5th IEEE European Conf. Antennas Propag.*, pp. 1684–1687, 2011.

[11] H. Zhao, J. Zhang, L. Yang, G. Pan, and M.-S. Alouini, "Secure mmWave communications in cognitive radio networks," *IEEE Wireless Commun. Lett.*, 2019.

[12] H. Hosseini, A. Anpalagan, K. Raahemifar, S. Erkucuk, and S. Habib, "Joint wavelet-based spectrum sensing and FBMC modulation for cognitive mmWave small cell networks," *IET Communications*, vol. 10, no. 14, pp. 1803–1809, 2016.

[13] S. Ulukus, A. Yener, E. Erkip, O. Simeone, M. Zorzi, P. Grover, and K. Huang, "Energy harvesting wireless communications: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, 2015.

[14] A. Kansal, J. Hsu, S. Zahedi, and M. B. Srivastava, "Power management in energy harvesting sensor networks," *ACM Trans. Embedded Comput. Sys.*, vol. 6, no. 4, p. 32, Sept. 2007.

[15] A. G. Marques, L. M. Lopez-Ramos, G. B. Giannakis, and J. Ramos, "Resource allocation for interweave and underlay CRs under probability-of-interference constraints," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 10, pp. 1922–1933, 2012.

[16] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[17] K. D. Huang, K. R. Duffy, and D. Malone, "H-RCA: 802.11 collision-aware rate control," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1021–1034, Aug. 2013.

[18] M. A. Qureshi and C. Tekin, "Fast learning for dynamic resource allocation in AI-enabled radio networks," *IEEE Trans. Cogn. Commun. Netw.*, pp. 1–1, 2019.

[19] S. Paladino, F. Trovò, M. Restelli, and N. Gatti, "Unimodal Thompson sampling for graph-structured arms," in *Proc. Conf. on Artif. Intell.*, pp. 2457–2463, Feb. 2017.

[20] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in *Proc. Int. Conf. Mach. Learn.*, pp. 521–529, Jan. 2014.

[21] C. Trinh, E. Kaufmann, C. Vernade, and R. Combes, "Solving Bernoulli rank-one bandits with unimodal Thompson sampling," *arXiv preprint arXiv:1912.03074*, 2019.

[22] S. Li, Z. Shao, and J. Huang, "ARM: Anonymous rating mechanism for discrete power control," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 326–340, 2016.

[23] C. Wu and D. P. Bertsekas, "Distributed power control algorithms for wireless networks," *IEEE Trans. Veh. Technol.*, vol. 50, no. 2, pp. 504–514, Mar. 2001.

[24] N. Devroye, M. Vu, and V. Tarokh, "Cognitive radio networks," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 12–23, 2008.

[25] K. T. Kim and S. K. Oh, "Cognitive ad-hoc networks under a cellular network with an interference temperature limit," in *Proc. 10th IEEE Int. Conf. Adv. Commun. Technol.*, vol. 2, pp. 879–882, 2008.

[26] I. Pardina Garcia, "Interference management in cognitive radio systems," Master's thesis, Universitat Politècnica de Catalunya, 2011.

[27] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Proc. Int. Conf. Algorithmic Learning Theory*, pp. 199–213, Springer, 2012.

[28] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in *Proc. Int. Conf. Artif. Intell. Statist.*, pp. 592–600, 2012.

[29] J. Komiyama, J. Honda, and H. Nakagawa, "Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays," in *Proc. Int. Conf. Mach. Learn.*, pp. 1152–1161, 2015.

[30] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, 2002.