2022 IEEE International Conference on Communications Workshops (ICC Workshops) | 978-1-6654-2671-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCWorkshops53468.2022.9882174

Adaptive Link Rate Selection for Throughput Maximization with Batched Thompson Sampling

Yuchao Chen[1,2], Haoyue Tang[3], Jintao Wang[1,2,4], Jian Song[1,2,4]
[1]Beijing National Research Center for Information Science and Technology (BNRist)
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China
[3]Department of Electronic Engineering, Yale University, New Haven, USA
[4]Research Institute of Tsinghua University in Shenzhen, Shenzhen, China
Email: cyc20@mails.tsinghua.edu.cn, tanghaoyue13@tsinghua.org.cn, {wangjintao; jsong}@tsinghua.edu.cn

***Abstract*—In this paper, we investigate the link rate selection for point-to-point throughput maximization under unknown channel statistics. The transmitter has limited policy switch times due to the constrained deployment opportunities. In order to maximize the expected system throughput, the wireless system must identify the optimal link rate as soon as possible. To meet this demand, we regard the time between two consecutive policy switching slots as a batch, and formulate the problem into the batched online sequential decision making framework. In particular, we propose two algorithms called the Modified/Constrained Batched Thompson Sampling (MBTS/CBTS), and show the expected regret is order-optimal with only logarithmic batches. Simulation results are provided to validate that the proposed algorithms can achieve a better regret and policy deployment trade-off compared with current state-of-the-art methods, i.e., greatly reducing the policy switch times with little regret growth.**

***Index Terms*—link rate selection, batched online learning, Thompson Sampling**

I. Introduction

Optimal link rate selection is one of the research hot spots in the wireless communication systems including the 802.11 systems [1] and cognitive radio networks [2]. In these scenarios, the system should identify the optimal rate among a finite set to maximize the expected throughput. Traditional link rate selection methods require reliable probing or channel estimation [3]. However, in the highly mobile setting, precise channel state information (CSI) feedback or even channel statistics may be unavailable. To overcome this challenge, online learning techniques, which only require historical observations for current decision making, can be used to guide selection strategies with only ACK/NACK outcomes to identify the optimal rate [4], [5].

Multi-armed bandit (MAB) problem, as a popular online sequential decision making problem, has been researched in-depth in recent years [6], [7]. Directly applying the MAB techniques to the link rate selection problem is widely studied in previous works [1], [8], [9]. In [1], the authors propose an Upper-Confidence-Bound (UCB) algorithm to achieve the order-optimal logarithmic expected regret. Later, Thompson Sampling (TS) algorithm [7] has also been modified to be applied to the rate selection problem in [8], [9]. In [8], the authors show the TS-based selection method can outperform the aforementioned UCB-based algorithm. Combining the prior information of the link rate, [9] further designs a constrained TS algorithm to improve the expected cumulative throughput performance.

However, the above algorithms require policy deployment in each time slot. This may be unpractical for many 802.11 devices with limited deployment opportunities or high policy switch cost. To overcome the challenge, batch learning is incorporated into the standard online learning framework to greatly reduce the deployment overhead. In the batched scheme, the transmitter has more freedom for controlling the policy deployment and only needs a subset of time slots for data computation and strategy switch.

Recent progress in batched online learning [10]–[12] has proven the achievability of the order-optimal logarithmic regret bound. In [10], the authors prove that it is sufficient to deploy $\mathcal{O}(\log T)$ batches to achieve the logarithmic regret over time horizon T, and apply a UCB-based successive elimination algorithm to achieve logarithmic regret. Similarly, batched TS algorithm is later designed in [12], and is also shown to outperform the above UCB-based methods empirically. However, these strategies require bounded [0,1] reward assumption and do not consider the prior information among arms. Therefore, new algorithms should be designed for this link rate selection problem.

In this paper, we consider a point-to-point discrete-time communication link, and aim to design a link rate selection strategy to maximize the expected system throughput. Due to the unknown channel statistics and limited policy switches, we resort to the batched online learning framework and propose a modified batched Thompson Sampling (MBTS) algorithm. We prove the proposed method can achieve the logarithmic regret bound with only logarithmic batch numbers. Moreover, with the knowledge of the prior information about the link rate, we propose a constrained batched Thompson Sampling (CBTS) algorithm to improve the throughput performance. Finally, simulation results are provided to validate the remarkable performance of the proposed algorithms compared with the current state-of-art methods.

This work was supported in part by Tsinghua University-China Mobile Research Institute Joint Innovation Center. *(Corresponding author: Jintao Wang)*

II. Problem Formulation

We consider a point-to-point discrete-time wireless communication link where a transmitter can transmit data at N different potential transmission rates, denoted by $r_1, r_2, \cdots, r_N$. Without loss of generality, we assume that $0 \leq r_1 < r_2 < \cdots < r_N$. For each transmission rate r_i, denote θ_i to be the probability of successful transmission at rate r_i. We assume that θ_i remains constant throughout the time horizon T. Since a successful transmission under a higher rate requires a better channel condition, we assume that the successful transmission probability θ_i is decreasing with respect to the rate index i, i.e., $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_N$. Denote $x_i(t) \in \{0, 1\}$ to record the data transmission outcome in slot t under rate r_i, which follows the Bernoulli distribution with parameter θ_i for all $t \in \{1, 2, \cdots, T\}$.

In order to maximize the expected system throughput in each slot t, the transmitter should identify the optimal link rate as soon as possible, i.e.,

$$i^\star = \arg\max_{i \in [N]} r_i \theta_i, \tag{1}$$

where $[N]$ denotes the set $\{1, 2, \cdots, N\}$ for convenience.

If the parameters θ_i are known accurately, the optimal link rate can be easily obtained by enumeration. Unfortunately, the probability θ_i is often unrevealed to the transmitter due to the unknown channel condition. Therefore, the transmitter must learn the parameters through historical observations and decisions during the communication process. In this case, the link rate selection problem can be formulated into the stochastic MAB problem. Each link rate r_i is the arm, and the transmission throughput in each slot $r_i x_i(t)$ can be viewed as the per-slot reward in the bandit problem. We use the expected regret $R(T)$ to measure the performance of any policy π over the time horizon T, which can be computed by

$$R(T) = \mathbb{E}_\pi \left[\sum_{t=1}^{T} (r_{i^\star} \theta_{i^\star} - r_{i(t)} \theta_{i(t)}) \right], \tag{2}$$

where $i(t)$ is the link rate selected in slot t under the policy π and the expectation is taken over the randomization of the policy π. Then, maximizing the expected system throughput is equivalent to minimizing the regret $R(T)$. In particular, the regret is expected to grow sub-linearly with time horizon T so that the average throughput is asymptotic optimal. However, the link rate selection problem has more prior information compared to the standard bandit problem, which arises in the dependence of the successful transmission probability, i.e., $\theta_i \geq \theta_j,\ \forall r_i < r_j$. This has been explained in detail in [8].

If the transmitter can update the selection strategy at will in each slot, recent progress in [8] and [9] proposes TS-based online learning algorithms to identify the optimal link rate through updating the posterior probability of θ_i. However, in a typical communication scenario, the transmitter may have limited policy switch times due to the high deployment costs. This motivates us to apply the batched online learning techniques to solve the link rate selection problem. To be specific, the transmitter can adaptively choose a number of time slots (much less than T) for strategy deployment and the transmission process between two consecutive policy switching time slots can be viewed as a batch. During a batch, the transmitter selects the link rate under a fixed policy; at the end of the batch, the transmitter can utilize the transmission decisions and outcomes during the batch to update the current selection strategy.

For the batched online learning problem, the transmitter should not only identify the optimal link rate quickly but also figure out the sweet point between a small number of batches and fast convergence to the optimal link rate. In the next section, we will propose several batched online learning algorithms to achieve the same order-optimal $\mathcal{O}(N \log T)$ regret performance in the standard stochastic MAB problem with just $\mathcal{O}(N \log T)$ batches.

III. Batched Thompson Sampling Algorithm

A. Algorithm Description

Batched Thompson Sampling (BTS) algorithm with adaptive batch selection has been proposed in [12] recently and is shown to outperform other algorithms including the UCB-based methods empirically. However, we cannot apply the BTS algorithm with Beta priors in [12] directly to our problem. This is because the reward distribution per slot $r_i x_i(t)$ in our problem is not supported on interval $[0, 1]$. In particular, the reward does not follow the Bernoulli distribution anymore. As an alternative, we propose the modified batched Thompson Sampling (MBTS) framework to design our link rate selection strategy, summarized in Algorithm 1.

Algorithm 1 Modified Batched Thompson Sampling (MBTS) Algorithm

Initialize: $n_i = l_i = 0,\ \alpha_i = \beta_i = S_i = F_i = 0,\ \forall i \in [N]$, batch = $\varnothing$.
for $t = 1, 2, \cdots, T$ **do**
  Sample each $\theta_i(t) \sim \text{Beta}(\alpha_i + 1, \beta_i + 1), \forall i \in [N]$.
  Transmit at rate $r_{i(t)}$, where

$$i(t) = \arg\max_{i \in [N]} r_i \theta_i(t). \tag{3}$$

  Update $n_{i(t)} = n_{i(t)} + 1$.
  Record $(S_{i(t)}, F_{i(t)}) = (S_{i(t)}, F_{i(t)}) + (x_{i(t)}(t), 1 - x_{i(t)}(t))$.
  if $n_{i(t)} < 2^{l_{i(t)}}$ **then**
    batch = batch $\bigcup \{i(t)\}$.
  else
    $l_{i(t)} = l_{i(t)} + 1$.
    Update the posterior probability parameters:

$$(\alpha_i, \beta_i) = (S_i, F_i),\ \forall i \in [N].$$

    Start a new batch: batch = $\varnothing$.
  end if
end for

The idea of the algorithm combines batch learning with the standard TS method. For each link rate r_i, the transmitter

keeps track of n_i, i.e., the number of times rate r_i is chosen. Variable l_i is defined to determine the batch length such that $2^{l_i-1} \leq n_i < 2^{l_i}$. Denote S_i and F_i to record the number of successful and failed transmissions up to the last time slot. Due to the limited policy switch times, the transmitter can only sample $\theta_i(t)$ with the information gathered until the last batch, i.e., α_i and β_i, instead of S_i and F_i. This is the main difference between the standard TS and the BTS method. Compared with the standard BTS with Beta priors proposed in [12], our MBTS algorithm computes the posterior probability of θ_i instead of the expected per-slot throughput $r_i\theta_i$. Despite the little modification, in the following part, we will show it will lead to remarkable regret performance compared with the general BTS-based method.

B. Performance Analysis

In this part, we will evaluate the performance of the MBTS algorithm, i.e., providing the upper bound for the number of batches and the expected regret. The proof is mainly based on [12] but needs some modification and improvement to be applied to our problem.

Theorem 1: [12, Theorem 4.1 Restated] The number of batches carried out by MBTS algorithm is $\mathcal{O}(N \log T)$.

Remark: Although previous works in [10] and [11] have proven that it is sufficient to deploy $\mathcal{O}(\log T)$ batches (independent of N) to achieve the order-optimal $\mathcal{O}(N \log T)$ regret in the batched stochastic MAB problem, the determination of batch size in these works depends on the time horizon T. However, in practice, the horizon T is always unknown. To overcome this challenge, the batch size carried out in the MBTS is adaptively chosen and is independent of the time horizon T.

It is a little more delicate to obtain the regret upper bound for the MBTS algorithm. For the following analysis, we mainly adopt the definition from [12], and reproduce here for convenience.

Definition 1 ($n_i(t)$ and $\hat{\mu}(t)$): Denote $n_i(t)$ to be the number of times rate r_i has been chosen until $t-1$. Define $\hat{\mu}_i(t)$ to be the empirical mean of the transmission outcomes x_i for rate r_i up to slot $t-1$.

Definition 2 ($\mathcal{F}_t$ and $B(t)$): For each time slot t, define the history of the selected rate and its outcomes as

$$\mathcal{F}_t := \{i(\tau), x_{i(\tau)}(\tau) | \tau \leq t\}.$$

Denote $B(t)$ to be the last time slot $t' \leq t-1$ such that the MBTS algorithm finishes a batch. Then, the information collected at the current slot t is the history $\mathcal{F}_{B(t)}$.

Definition 3 (Thresholds x_i and y_i): For each sub-optimal link rate $r_i (i \neq i^\star)$, we choose two thresholds x_i and y_i (determined in later analysis) such that $r_i\theta_i < r_i x_i < r_i y_i < r_{i^\star}\theta_{i^\star}$.

Definition 4 (Events $E_i^\theta(t)$ and $E_i^\mu(t)$): Define $E_i^\theta(t)$ to be the event $\{\theta_i(t) \leq y_i\}$, and $E_i^\mu(t)$ to be the event $\{\hat{\mu}_i(B(t)) \leq x_i\}$.

Definition 5 (Probability $p_{i,t}$): Define the probability $p_{i,t}$ as $p_{i,t} = \mathbb{P}(r_{i^\star}\theta_{i^\star}(t) > r_i y_i | \mathcal{F}_{B(t)}) = \mathbb{P}(\theta_{i^\star}(t) > \frac{r_i y_i}{r_{i^\star}} | \mathcal{F}_{B(t)})$.

The following theorem provides the problem-dependent regret upper bound of the MBTS algorithm:

Theorem 2: The expected regret until the time horizon T for the MBTS algorithm can be upper bounded by

$$R(T) \leq (1+\epsilon) \sum_{i \neq i^\star} \frac{\mathbb{I}\left(\frac{r_{i^\star}\theta_{i^\star}}{r_i} \leq 1\right) \log T}{D\left(\theta_i, \frac{r_{i^\star}\theta_{i^\star}}{r_i}\right)} \Delta_i + \mathcal{O}\left(\frac{N}{\epsilon^2}\right), \quad (4)$$

where $\epsilon \in (0, 1]$, $\Delta_i = r_{i^\star}\theta_{i^\star} - r_i\theta_i$, and $D(a, b)$ is the KL divergence between two Bernoulli distributions with parameters a and b, respectively.

Proof: The proof is provided in Appendix A. ∎

C. General Batched Thompson Sampling Algorithm

Although the BTS algorithm proposed in [12] is mainly analyzed with Beta and Gaussian priors, [7] has provided a straightforward way to generalize the TS algorithm to the general priors. Therefore, we can also design a general batched Thompson Sampling (GBTS) algorithm for our link rate selection problem. The idea of GBTS is to normalize the transmission throughput such that $y(t) = \frac{r_{i(t)}}{r_N} x_{i(t)}(t)$. The detailed description is omitted due to the space limitation. The following theorem is an immediate result of Theorem 4.3 in [12] for the GBTS algorithm.

Theorem 3: The expected regret of GBTS algorithm until time T can be upper bounded by

$$R(T) \leq (1+\epsilon) \sum_{i \neq i^\star} \frac{\log T}{D\left(\frac{r_i}{r_N}\theta_i, \frac{r_{i^\star}}{r_N}\theta_{i^\star}\right)} \Delta_i + \mathcal{O}\left(\frac{N}{\epsilon^2}\right), \quad (5)$$

where $\epsilon \in (0, 1]$.

Comparing the regret bound in Theorem 2 and Theorem 3, we will find our MBTS algorithm is expected to outperform the GBTS method since for those parameters such that $\frac{r_{i^\star}\theta_{i^\star}}{r_i} > 1$, MBTS can achieve $\mathcal{O}(1)$ regret (Case 2). This implies that MBTS can quickly distinguish the link rate whose expected throughput $r_i\theta_i$ is far from optimality.

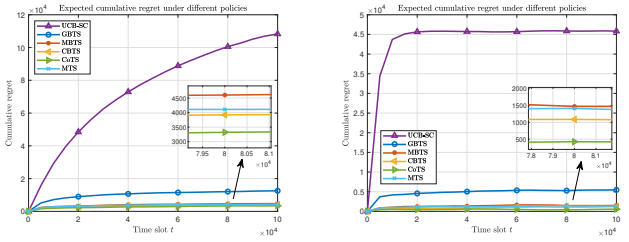
D. Constrained Batched Thompson Sampling Algorithm

As mentioned in the previous section, the successful transmission probability θ_i depends on each other, i.e., $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_N$. However, both MBTS and GBTS algorithms do not utilize this prior information explicitly. Therefore, we follow the instruction of [9] to improve the algorithm performance by combining the prior information, and propose the constrained batched Thompson Sampling (CBTS) algorithm. The idea of CBTS is to sample a feasible θ_i satisfying the constraint. Denote $\boldsymbol{\theta}(t) = (\theta_1(t), \theta_2(t), \cdots, \theta_N(t))$ to be the sample vector in slot t, and Θ to be the feasible set satisfying the prior information, i.e.,

$$\Theta = \{(\theta_1, \theta_2, \cdots, \theta_N) | \theta_1 \geq \theta_2 \geq \cdots \geq \theta_N\}. \quad (6)$$

Based on MBTS described in Algorithm 1, CBTS improves the step 3 to ensure the sample vector $\boldsymbol{\theta}(t)$ in each slot belongs to the feasible set Θ, i.e.,

$$\boldsymbol{\theta}(t) \sim \mathbb{I}(\boldsymbol{\theta}(t) \in \Theta) \prod_{i=1}^{N} \text{Beta}(\alpha_i + 1, \beta_i + 1). \quad (7)$$



(a) gradual case (b) steep case

Fig. 1. Expected regret performance in two cases.

In practice, this step can be done through discarding all the infeasible $\boldsymbol{\theta}(t)$ until sampling a feasible one, called "rejection sampling" in [9].

IV. Simulation Results

In this section, we provide simulation results to validate the performance of our proposed algorithms. We consider a single-link 802.11g system with $N = 8$ possible link rates (same in [1]), i.e., $[6, 9, 12, 18, 24, 36, 48, 54]$ (in Mbit/s). We evaluate the performance of our proposed algorithms under different successful transmission probability θ_i (same in [9]) in $T = 10^5$ consecutive slots, where each algorithm is run 100 times for average.

First, we consider the gradual θ_i case, i.e., the corresponding successful transmission probability for each r_i is $[0.95, 0.9, 0.8, 0.65, 0.45, 0.25, 0.15, 0.1]$. Through computation, the expected throughput for each link rate $r_i\theta_i$ is $[5.7, 8.1, 9.6, 11.7, 10.8, 9, 7.2, 5.4]$. Therefore, the optimal link rate $r_{i^\star} = 18$Mbit/s. We validate the performance of the proposed MBTS, CBTS, and GBTS algorithms compared with the UCB-based successive elimination (UCB-SC) algorithm in [11] and other current link rate selection policies without batch learning such as Modified TS (MTS) in [8], and Constrained TS (CoTS) in [9], as depicted in Fig. 1(a).

For consistency, we set the batch numbers $B = \lceil N \log T \rceil$ in the UCB-SC algorithm, where $\lceil \cdot \rceil$ is the ceiling function. Although all the algorithms depicted in Fig. 1(a) are proven to achieve the $\mathcal{O}(N \log T)$ regret performance theoretically, our proposed MBTS and CBTS algorithms outperform the other batched link rate selection policies. In general, the TS-based algorithms can achieve better performance than the UCB-based method. This is because the TS-based algorithm selects the link rate according to the posterior probability, while the UCB-based algorithm selects the candidate of all possible optimal link rates uniformly. Therefore, the sub-optimal link rate will be selected more times under the UCB-SC, especially in the early period. The superiority of MBTS and CBTS has been explained in the above section, i.e., the proposed algorithms can achieve $\mathcal{O}(1)$ regret for the distinct sub-optimal link rate. To better show this advantage, Table I records the selection times for all the link rates under different policies in one sample path. Notice that rate r_1 and r_2 satisfy the condition $\frac{r_{i^\star}\theta_{i^\star}}{r_i} > 1$. Therefore, both MBTS and CBTS discard these two options quickly even without choosing them, which leads to better regret performance.

Compared with the link rate selection policies without batch learning, it is intuitive to expect the high regret of batch learning algorithm at the cost of low policy switching times. In fact, both CoTS and MTS methods require $T = 10^5$ times of policy switch, while our proposed algorithms only need at most 132 times of updating policy. However, Fig. 1(a) demonstrates that our CBTS algorithm can still outperform the MTS method with the help of the prior information about the link rate.

TABLE I
Selection times for different link rate in the gradual case

Policy	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
UCB-SC	1967	4159	12730	39885	38850	1349	631	429
GBTS	158	393	143	97575	977	492	128	134
MBTS	0	0	85	96990	2235	316	235	139
CBTS	0	0	5	96920	2217	663	172	23

Next, we consider the steep θ_i case, i.e., the transmission probability is either very high or quite low. We choose θ_i as $[0.99, 0.98, 0.96, 0.93, 0.9, 0.1, 0.06, 0.04]$, and the expected throughput can be computed as $[5.94, 8.82, 11.52, 16.74, 21.6, 3.6, 2.88, 2.16]$. The expected and single sample path regret performance are depicted in Fig. 1(b) and Table II, respectively. As portrayed in Fig. 1(b), CBTS and MBTS can also achieve better performance than the other two batch selection policies, and perform competitively compared with MTS and CoTS without batch learning. This again demonstrates that our proposed algorithms can achieve a better trade-off between low regret and low policy switch times.

However, while CBTS can outperform MBTS by considering the prior information of θ_i, it increases the computational complexity significantly with the rejection sampling method. This is because the probability of sampling a feasible $\boldsymbol{\theta}(t)$ is quite small when θ_5 is close to 1, and θ_1 to θ_4 are almost sampled from a uniform distribution between $[0, 1]$ (since they are hardly chosen). In fact, since $\frac{r_{i^\star}\theta_{i^\star}}{r_i}$ is quite larger than 1 when i is small in this case, MBTS algorithm can also quickly figure out the optimal link rate. Therefore, the rejection sampling method is not always an efficient way to carry out the CBTS algorithm, and MBTS is satisfactory enough to achieve the trade-off between the low regret and the computational complexity.

TABLE II
Selection times for different link rate in the steep case

Policy	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
UCB-SC	377	488	924	3242	94380	290	129	170
GBTS	46	70	140	189	99486	26	21	22
MBTS	0	0	0	4	99932	18	25	21
CBTS	0	0	0	2	99966	19	11	2

V. Conclusion

In this paper, we propose two algorithms called MBTS and CBTS to maximize the expected cumulative throughput in the point-to-point rate selection problem. These proposed methods can adaptively identify the optimal link rate with unknown channel statistics, and are proven to achieve the order-optimal logarithmic regret bound with only logarithmic policy switch times. Through simulation results, we validate the power of batch learning in this problem, which can reduce the policy deployment cost with little performance deterioration. Interesting extensions include the rate selection strategy in the multi-user scenario. The challenge mainly comes from the joint batch length determination and the corresponding theoretical convergence proof, which will be our future work.

Appendix A
Proof of Theorem 2

Notice that the expected regret defined in Eq. (2) can be rewritten as:

$$R(T)=\sum_{i\neq i^\star}\Delta_i\mathbb{E}[n_i(T+1)]. \tag{8}$$

Therefore, it is sufficient to upper bound $\mathbb{E}[n_i(T+1)]$ for each $i\neq i^\star$. As in [12], we split $\mathbb{E}[n_i(T+1)]$ into three terms based on the events defined in Definition 4:

$$\begin{aligned}\mathbb{E}[n_i(T+1)]&=\sum_{t=1}^{T}\mathbb{P}(i(t)=i)\\&=\sum_{t=1}^{T}\mathbb{P}(i(t)=i,E_i^\mu(t),E_i^\theta(t))\\&\quad+\sum_{t=1}^{T}\mathbb{P}(i(t)=i,E_i^\mu(t),\overline{E_i^\theta(t)})\\&\quad+\sum_{t=1}^{T}\mathbb{P}(i(t)=i,\overline{E_i^\mu(t)}),\end{aligned} \tag{9}$$

where event $\bar{A}$ denotes the complement of event A.

For the first term in Eq. (9), we claim that it can be upper bounded by $\mathcal{O}(1)$, whose proof is provided in Appendix B.

Now consider the second term in Eq. (9). Different from [12], we study the following two cases:

Case 1: $\frac{r_{i^\star}\theta_{i^\star}}{r_i}\leq 1$.

In this case, we have $\theta_i<x_i<y_i<\frac{r_{i^\star}\theta_{i^\star}}{r_i}\leq 1$. Applying [12, Lemma A.14], we have:

$$\sum_{t=1}^{T}\mathbb{P}(i(t)=i,E_i^\mu(t),\overline{E_i^\theta(t)})\leq L_i(T)+1, \tag{10}$$

where $L_i(T)=\frac{\log T}{D(x_i,y_i)}$.

Case 2: $\frac{r_{i^\star}\theta_{i^\star}}{r_i}>1$.

In this case, we have $\frac{r_i\theta_i}{r_{i^\star}}<\frac{r_ix_i}{r_{i^\star}}<\frac{r_iy_i}{r_{i^\star}}<\theta_{i^\star}\leq 1$. Then, we can choose $y_i\in(1,\frac{r_{i^\star}\theta_{i^\star}}{r_i})$ such that the event $\mathbb{P}(\overline{E_i^\theta(t)})=\mathbb{P}(\theta_i(t)>y_i)=0$. Therefore, we have

$$\sum_{t=1}^{T}\mathbb{P}(i(t)=i,E_i^\mu(t),\overline{E_i^\theta(t)})=0, \tag{11}$$

Combining the above two cases, we can upper bound the second term in Eq. (9) as follows

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{P}(i(t)=i,E_i^\mu(t),\overline{E_i^\theta(t)})\\&\leq\mathbb{I}\left(\frac{r_{i^\star}\theta_{i^\star}}{r_i}\leq 1\right)\left(\frac{\log T}{D(x_i,y_i)}+1\right).\end{aligned} \tag{12}$$

For the last term in Eq. (9), we directly apply [12, Lemma A.13] to have

$$\sum_{t=1}^{T}\mathbb{P}(i(t)=i,\overline{E_i^\mu(t)})\leq\frac{2}{D(x_i,\theta_i)}+1. \tag{13}$$

Now it remains to determine the threshold x_i and y_i when $\frac{r_{i^\star}\theta_{i^\star}}{r_i}\leq 1$. To obtain the problem-dependent regret bound, for $0\leq\epsilon<1$, we choose x_i such that $D(x_i,\frac{r_{i^\star}\theta_{i^\star}}{r_i})=\frac{D(\theta_i,\frac{r_{i^\star}\theta_{i^\star}}{r_i})}{1+\epsilon}$, and y_i such that $D(x_i,y_i)=\frac{D(x_i,\frac{r_{i^\star}\theta_{i^\star}}{r_i})}{1+\epsilon}=\frac{D(\theta_i,\frac{r_{i^\star}\theta_{i^\star}}{r_i})}{(1+\epsilon)^2}$. After some manipulations, we have:

$$x_i-\theta_i\geq\frac{\epsilon}{1+\epsilon}\frac{D(\theta_i,\frac{r_{i^\star}\theta_{i^\star}}{r_i})}{\log\left(\frac{r_{i^\star}\theta_{i^\star}(1-\theta_i)}{\theta_i(r_i-r_{i^\star}\theta_{i^\star})}\right)}.$$

Then the Pinsker's inequality implies that $\frac{1}{D(x_i,\theta_i)}\leq\frac{1}{2(x_i-\theta_i)^2}=\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$. Combining Eq. (17), Eq. (12) and Eq. (13) with the choice of x_i and y_i, we have

$$\begin{aligned}&\mathbb{E}[n_i(T+1)]\\&\leq\mathcal{O}(1)+\mathbb{I}\left(\frac{r_{i^\star}\theta_{i^\star}}{r_i}\leq 1\right)(1+\epsilon)^2\frac{\log T}{D(\theta_i,\frac{r_{i^\star}\theta_{i^\star}}{r_i})}+\mathcal{O}(\frac{1}{\epsilon^2})\\&\leq(1+\epsilon')\frac{\mathbb{I}\left(\frac{r_{i^\star}\theta_{i^\star}}{r_i}\leq 1\right)\log T}{D(\theta_i,\frac{r_{i^\star}\theta_{i^\star}}{r_i})}+\mathcal{O}(\frac{1}{\epsilon'^2}),\end{aligned}$$

where $\epsilon'=3\epsilon$. Combining the result with Eq. (8) yields Theorem 2.

Appendix B
Proof of the $\mathcal{O}(1)$ bound of the first term in Eq. (9)

First, we establish a relationship between the probability of selecting any sub-optimal rate r_i and the optimal $r_{i^\star}$ while two events $E_i^\theta(t)$ and $E_i^\mu(t)$ happen.

Lemma 1: For all rate $r_i,\ i\neq i^\star$, we have

$$\begin{aligned}&\mathbb{P}(i(t)=i,E_i^\mu(t),E_i^\theta(t)|\mathcal{F}_{B(t)})\\&\leq\frac{1-p_{i,t}}{p_{i,t}}\mathbb{P}(i(t)=i^\star,E_i^\mu(t),E_i^\theta(t)|\mathcal{F}_{B(t)}).\end{aligned} \tag{14}$$

Proof: The proof is similar to the one in [8, Lemma 1], and thus is omitted here. ■

With Lemma 1, we can bound the first term in Eq. (9) as

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{P}(i(t)=i,E_i^{\mu}(t),E_i^{\theta}(t))\\
\leq&\sum_{t=1}^{T}\mathbb{E}[\mathbb{P}(i(t)=i,E_i^{\mu}(t),E_i^{\theta}(t)|\mathcal{F}_{B(t)})]\\
\leq&\sum_{t=1}^{T}\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\mathbb{P}(i(t)=i^{\star},E_i^{\mu}(t),E_i^{\theta}(t)|\mathcal{F}_{B(t)})\right]\\
=&\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\mathbb{I}(i(t)=i^{\star},E_i^{\mu}(t),E_i^{\theta}(t)|\mathcal{F}_{B(t)})\right]\right]\\
\leq&\sum_{t=1}^{T}\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\mathbb{I}(i(t)=i^{\star},E_i^{\mu}(t),E_i^{\theta}(t))\right], \qquad (15)
\end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Here we correct the notation and analysis in [12]. Denote τ_k' to be the time step that rate $r_{i^\star}$ has been selected for the k-th time ($\tau_0'=0$). Then, we define another sequence $\{\tau_k\}$ such that

$$\tau_k=\begin{cases}B(\tau_{k+1}'), & \text{if } \tau_k' \text{ is the last selection of } r_{i^\star}\\ & \text{in the current batch;}\\ \tau_k', & \text{else.}\end{cases}$$

Then we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\mathbb{I}(i(t)=i^{\star},E_i^{\mu}(t),E_i^{\theta}(t))\right]\\
\overset{(a)}{\leq}&\sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\sum_{t=\tau_k+1}^{\tau_{k+1}}\mathbb{I}(i(t)=i^{\star},E_i^{\mu}(t),E_i^{\theta}(t))\right]\\
\overset{(b)}{\leq}&\sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\right], \qquad (16)
\end{aligned}$$

where (a) holds since we divide the whole time horizon by τ_k and $p_{i,t}$ remains the same in each interval, and (b) holds since in $t\in[\tau_k+1,\tau_{k+1}]$, $r_{i^\star}$ is selected at most once by the definition of τ_k.

Then, similar to [7, Lemma 2.9], we can upper bound the term $\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\right]$ as follows:

Lemma 2: For any sub-optimal rate $r_i,\ i\neq i^\star$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}-1\right]\\
\leq&\begin{cases}\frac{3}{\Delta_i'}, \text{for } n_{i^\star}(B(\tau_k+1))<\frac{8}{\Delta_i'^2},\\ \Theta\left(e^{-\frac{\Delta_i'^2k}{4}}+\frac{e^{-\frac{D_ik}{2}}}{\left(\frac{k}{2}+1\right)\Delta_i'^2}+\frac{1}{e^{\frac{\Delta_i'^2k}{16}}-1}\right), \text{else,}\end{cases}
\end{aligned}$$

where $\Delta_i'=\theta_{i^\star}-\frac{r_iy_i}{r_{i^\star}}$, and $D_i=D\left(\frac{r_iy_i}{r_{i^\star}},\theta_{i^\star}\right)$.

Remark: We correct the mistake made in [12, Lemma A.11], although it does not affect the following analysis in that paper. The above lemma can be obtained by replacing k in [7, Lemma 2.9] with $n_{i^\star}(B(\tau_k+1))$ and applying $n_i(B(t))\geq\frac{1}{2}n_i(t)$ deduced in [12, Lemma A.9].

Combining Eq. (15), Eq. (16) and Lemma 2, we can upper bound the first term in Eq. (9) as

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{P}(i(t)=i,E_i^{\mu}(t),E_i^{\theta}(t))\\
\leq&\frac{3}{\Delta_i'}\mathbb{I}\left(n_{i^\star}(B(\tau_k+1))<\frac{8}{\Delta_i'^2}\right)\\
&+\sum_{k=0}^{T-1}\Theta\left(e^{-\frac{\Delta_i'^2k}{4}}+\frac{e^{-\frac{D_ik}{2}}}{\left(\frac{k}{2}+1\right)\Delta_i'^2}+\frac{1}{e^{\frac{\Delta_i'^2k}{16}}-1}\right)\\
\overset{(a)}{\leq}&\frac{3}{\Delta_i'}\mathbb{I}\left(k<\frac{16}{\Delta_i'^2}\right)\\
&+\sum_{k=0}^{T-1}\Theta\left(e^{-\frac{\Delta_i'^2k}{4}}+\frac{e^{-\frac{D_ik}{2}}}{\left(\frac{k}{2}+1\right)\Delta_i'^2}+\frac{1}{e^{\frac{\Delta_i'^2k}{16}}-1}\right)\\
\leq&\frac{48}{\Delta_i'^2}+\Theta\left(\frac{1}{\Delta_i'^2}+\frac{1}{D_i\Delta_i'^2}+\frac{1}{\Delta_i'^4}\right)=\mathcal{O}(1). \qquad (17)
\end{aligned}$$

where (a) holds by [12, Lemma A.9].

REFERENCES

[1] R. Combes, A. Proutiere, D. Yun, J. Ok, and Y. Yi, "Optimal Rate Sampling in 802.11 systems," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014, pp. 2760–2767.

[2] R. Combes and A. Proutiere, "Dynamic Rate and Channel Selection in Cognitive Radio Systems," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 5, pp. 910–921, 2015.

[3] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[4] R. Aggarwal, P. Schniter, and C. E. Koksal, "Rate Adaptation via Link-Layer Feedback for Goodput Maximization over a Time-Varying Channel," *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4276–4285, 2009.

[5] C. E. Koksal and P. Schniter, "Robust Rate-Adaptive Wireless Communication Using ACK/NAK-Feedback," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1752–1765, 2012.

[6] 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. [Online]. Available: https://doi.org/10.1023/A:1013689704352

[7] S. Agrawal and N. Goyal, "Near-Optimal Regret Bounds for Thompson Sampling," *J. ACM*, vol. 64, no. 5, pp. 30:1–30:24, 2017. [Online]. Available: https://doi.org/10.1145/3088510

[8] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-Complexity, Low-Regret Link Rate Selection in Rapidly-Varying Wireless Channels," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 540–548.

[9] ——, "Link Rate Selection using Constrained Thompson Sampling," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019, pp. 739–747.

[10] Z. Gao, Y. Han, Z. Ren, and Z. Zhou, "Batched Multi-armed Bandits Problem," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[11] H. Esfandiari, A. Karbasi, A. Mehrabian, and V. S. Mirrokni, "Regret Bounds for Batched Bandits," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI Press, 2021, pp. 7340–7348. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16901

[12] A. Karbasi, V. S. Mirrokni, and M. Shadravan, "Parallelizing Thompson Sampling," *CoRR*, vol. abs/2106.01420, 2021. [Online]. Available: https://arxiv.org/abs/2106.01420