2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) | 978-1-6654-6901-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/BMSB55706.2022.9828660

Information Directed Learning Algorithm for Minimizing Queue Length Regret

Xinyu Hou, Haoyue Tang, Jintao Wang and Jian Song
Department of Electronic Engineering, Tsinghua University
Beijing National Research Center for Information Science and Engineering, Beijing, China
Research Institute of Tsinghua University in Shenzhen, Shenzhen, China
{houxy19@mails, thy17@mails, wangjintao@, jsong@}.tsinghua.edu.cn

***Abstract*—In this paper, we consider a discrete time transmitter-receiver pair with K error-prone transmission channels. In each slot, packets arrive at the transmitter randomly and wait in the queue before they are successfully delivered to the receiver. The goal is to design an adaptive channel selection strategy to minimize the queue length over T consecutive slots in the absence of packet-loss probabilities. We categorize the current queueing status into two classes: (1) when the queue is empty, we fully explore all the channels through uniform sampling; (2) when there are untransmitted packets in the queue, we balance the explore-exploit trade-off using information directed sampling. We prove that the proposed algorithm reaches a time cumulative queue length regret of order $\mathcal{O}(1)$. Simulation results validate the effectiveness of the proposed algorithm.**

I. Introduction

High Dynamic Range (HDR) and Ultra High Definition (UHD) video streaming, 4K live broadcasting applications require communications to be ultra high reliable and low latency [1]. Due to the limited communication resources, the transmitter needs to figure out the best transmission strategy. However, the statistical information (e.g., packet-loss probabilities of each channel) is hard to obtain and it is important to design online algorithm that identifies the best transmission strategy as soon as possible.

Channel and link rate selection in the absence of channel statistical information has been an important research problem over the years. The well known Max-Weight strategy [2] stabilizes the queue length for each node in multi-hop networks when the arrival rate of the packets is within the stability region of the system [3]. However, it incurs high transmission delay when the total throughput of the entire network approaches the stability region. To reduce the transmission delay, [4] combined learning algorithm with the Max-Weight policy for BS activation and Huang *et al.* proposed an online scheduling algorithm through dual learning [5]. Although those proposed algorithms can effectively minimize the average queue length compared with the Max-Weight algorithm, the theoretic behaviour of the queue length regret is not fully understood.

The multi-arm bandit algorithms (MAB) provide efficient solutions to online sequential decision making problems with theoretic guarantees [6]. The cost of unknown statistical information is characterized through the cumulative regret, i.e., the difference between the expected cumulative cost of the proposed algorithm and optimum algorithm when the statistical information is known. Lai and Robbins showed in [7] that the regret of classic MAB grows at least with order $\mathcal{O}(\log T)$ with time T and the UCB/Thompson sampling algorithms have been shown to approach this convergence bound. Notice that the convergence result from [7] is established under the condition that cost occurs in every slot, while in a queuing system, no cost is incurred when the queue is empty and the convergence bound is $\mathcal{O}(1)$. This calls for more dedicated design for online channel selection for approaching the converse bound.

Recently, [8] proposes an online channel selection algorithm by dividing time slots into busy and empty periods based on the queuing states of the system. When the queue is empty, the transmitter fully explores the channel states by uniformly selecting each channel and UCB based algorithm is used to balance the explore-exploit trade-off during the busy period. Notice that both UCB and Thompson sampling algorithms perform poorly when the prior distributions of the bandits are complex. In this paper, we aim at improving the online channel selection algorithm from [8] by utilizing the information directed sampling algorithm (IDS) [9] during the busy periods. By quantifying the benefit and cost of exploration as the mutual information gain and the difference of empirical reward respectively, the average cost-per-bit of choosing a sub-optimal action can be defined as the mutual information gain divided by the difference of the empirical reward. The proposed IDS algorithm thus optimizes the explore-exploit trade-off by minimizing the cost-per-bit information and therefore achieves a smaller cumulative queue length regret.

The rest of this paper is organized as follows: We present the system model and formulate the optimization problem in Section II. Section III introduces our designed Uniformly Exploration and Information Directed Stabilizing Algorithm (UE-IDS). The performance of the proposed algorithm is validated through numerical simulations in Section IV and Section V draws the conclusion.

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

II. Problem Formulation

We consider a transmitter sending information packets to the receiver via K erroneous channels, as depicted in Fig. 1. Let the time be slotted, and the index of the current slot is denoted by $t = 0, 1, \cdots, T$. At the beginning of each slot, packets arrive randomly at the transmitter. Let $A(t) = 1$ be the identification a packet has arrived in slot t; otherwise $A(t) = 0$. We assume each $A(t)$ is i.i.d and follows the Bernoulli distribution with $\mathbb{E}[A(t)] = \lambda, \lambda \in (0, 1]$. In each slot t, due to the wireless interference constraint, the transmitter can select only one of K channels for sending out packets. We assume each transmission attempt takes one slot and if the transmitted packet is successfully received, we denote $X(t) = 1$, and an ACK will be received at the transmitter by the end of slot t. Let $i(t) \in [K]$ be the index of the selected channel. Note that the quality of each channel is different, we assume when $i(t) = i$, the transmission outcome $X(t) = X_i(t)$ satisfies Bernoulli process with $\mathbb{E}[X(t)] = \mu_i$, and $X(t)$ is independent of $X(t')$ in other slots.

Assumption 1: We assume that the arrival rate is within the stability region of the system. At least one service rate is larger than the arrival rate, i.e., $\mu_{i^*} > \lambda$.

We assume the untransmitted packets wait in the queue at the transmitter, and let $Q(t)$ be the queue-length at the beginning of slot t. Notice that $X(t)$ is also the number of transmitted packets in slot t, then the evolution of the queue length $Q(t)$ is as follows:

$$Q(t+1) = (Q(t) - X(t))^+ + A(t), \tag{1}$$

where $(\cdot)^+ = \max\{0, \cdot\}$. We assume that when the queue is empty, i.e., $Q(t) = 0$, the transmitter can also select a channel to send testing packets and obtain feedback $X(t)$ of channel $i(t)$.

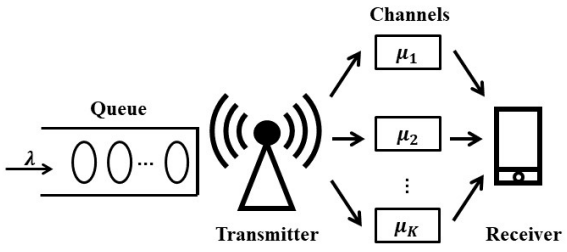

Fig. 1. System Model

Assume that the transmission statistics $\{\mu_i\}_{i=1}^K$ is unknown to the transmitter, the goal of our research is to minimize the expected cumulative queue length by designing a transmission strategy $\pi(t)$ in each slot t using the historical transmission feedback $\mathcal{F}_{t-1} := \sigma(\{(i(0), X(0)), \cdots, (i(t-1), X(t-1))\})$, where $\sigma(\cdot)$ is the $\sigma-$algebra generated by random variables. Let $Q^\pi(t)$ be the queue length by using scheduling policy π in slot t. The problem is organized as follows:

Problem:

$$\min_\pi \mathbb{E}\left[\sum_{t=0}^{T-1} Q^\pi(t)\right], \quad \text{where } \pi : \mathcal{F}_{t-1} \to i(t). \tag{2}$$

Notice that when $\{\mu_i\}_{i=1}^K$ is known, the optimum policy π^* that minimizes the queue length is $\pi^*(t) \equiv i^*$, where the index of the best channel is denoted by $i^* \triangleq \arg\max_{i\in[K]} \mu_i$.[1] To evaluate the performance of a scheduling policy π, we compare the expected queue length using policy π with the expected queue length using the optimum policy π^*. By definition, the cumulative queue length regret applying scheduling policy π at slot T is denoted by $R^\pi(T)$, which can be computed by:

$$R^\pi(T) \triangleq \mathbb{E}\left[\sum_{t=0}^{T-1} Q^\pi(t) - \sum_{t=0}^{T-1} Q^*(t)\right]. \tag{3}$$

III. Learning Algorithm Design

The above channel selection problem in the absence of channel statistics $\{\mu_i\}_{i=1}^J$ can be cast into the multi-arm bandit framework, where each channel $i \in [K]$ can be viewed as an arm and the queue length $Q(t)$ is the cost in slot t. An efficient bandit algorithm achieves a trade-off between exploration and exploitation. Unlike a traditional sequential decision making problem where a reward/cost is received in every slot, in our problem, when $Q(t) = 0$, no cost is incurred. Therefore, using full exploration (i.e., do not choose the channel with the highest empirical success rate) when $Q(t) = 0$ does not increase the total cost. Our algorithm improves the queue length performance by doing full exploration for slots with $Q(t) = 0$ (the so called "*idle* period") and improves the explore-exploit trade-off for slots with $Q(t) > 0$ (referred as "*busy* period") using the Information Directed sampling (IDS) [9]. The detailed algorithm design is as follows:

A. Queue State Classification

a) Idle Period (Complete Exploration): In the idle period $Q(t) = 0$, we perform channel exploration by selecting a channel $i \in [K]$ uniformly and randomly. Let $s_i(t)$ and $f_i(t)$ be the total successful and failed transmission times of arm i up to slot t. We then update $s_i(t)$ and $f_i(t)$ based on the transmission feedback $X(t)$ in slot t, i.e., $s_i(t) = s_i(t-1) + X_i(t)$, $f_i(t) = f_i(t-1) + 1 - X_i(t)$. Then the experimental mean reward of each channel is $\hat{\mu_i}(t) = \frac{s_i(t)}{s_i(t)+f_i(t)}$. The decision of choosing which channel in the idle period doesn't effect the queue's backlog, hence the cumulative queue regret is not growing.

b) Busy Period (Improving Explore-Exploit Trade-off using IDS): Recall that $Q(t)$ is the length of the queue in slot t and $Q(t) > 0$ means the queue is non-empty. Let τ_d be the d-th starting time when the queue leaves the empty state, i.e., $\tau_1 := \min_t\{Q(t) > 0\}$ and $\tau_d := \min_{t>\tau_{d-1}}\{Q(t) > 0\}$. The

[1]For simplicity, assume there is only one best channel with the highest transmission success probability.

channel selection algorithm during each busy period d can be divided into two parts:

- If the busy period has lasted less than q_d slots, we do full exploitation by choosing channel with the highest empirical success rate, i.e., $i(t) := \arg\max_{i\in[K]} \frac{s_i(t)}{s_i(t)+f_i(t)}$.
- If the busy period lasts for more than q_d slots, we achieve the explore-exploit trade-off using IDS. Then the information gain $g_t(i)$ of choosing arm i can be computed by:

$$g_t(i) := \mathbb{E}[H(i^*|\mathcal{F}_{t-1}) - H(i^*|\mathcal{F}_{t-1}, i(t) = i)], \quad (4)$$

where $H(i^*|\mathcal{F}_t)$ is the conditional entropy of random variable α_t. The posterior distribution of μ_i given $s_i(t)$ and $f_i(t)$ follows a Beta distribution, i.e. $\Pr(\mu_i = \theta_i) = \text{Beta}(\theta_i; s_i(t), f_i(t))$. The expected reward of the best channel given past observations is denoted by $\rho^*(t) := \mathbb{E}[\max_i \mu_i|\mathcal{F}_t]$. Let $\Delta_i(t) := \rho^*(t) - \hat{\mu}_i$ be the difference between $\theta^*(t)$ and the empirical mean of arm i. Define the information ratio as $\frac{\Delta_i(t)^2}{g_i(t)}$. The schedule policy in the $(d+1)$-th slot in busy period d after slot $\tau_d + q_d$ is to select channel i with probability $p_i(t)$, where $p_i(t)$ can be computed by:

$$p_i(t) = \min_{p\in C_K} \frac{(p^\top \vec{\Delta}(t))^2}{p^\top \vec{g}(t)}. \quad (5)$$

Denote $C_K = \{p \in \mathbb{R}_+^K : \sum_i p_i = 1\}$ as the K-dimensional unit simplex. Computations of $\Delta_i(t)$ and $g_i(t)$ can be found in Algorithm 2, and the detailed derivations are provided in the appendix.

B. Algorithm Design

The overall algorithm contains two parts: exploration in idle period and exploitation in busy period. The algorithm flow chart is shown in Algorithm 1. At first, the algorithm sets empty record of transmission feedback and empty queue backlog, the index of busy period is 0. As the system knows the number of arrival packets and served packets, the queue length at each slot can be calculated, then the schedule policy will operate depending on whether the queue is empty or not.

- **For idle period:** The schedule policy is uniformly and randomly selecting channels at each slot in order to update each arm's empirical mean reward. Since the queue backlog is empty, the transmission result won't affect queue length at the next slot.
- **For busy period:** The index of busy period d grows when comes to a busy period. Let l be the l-th slot in this busy period. If $l \le q_d$, the schedule policy is selecting the channel with the highest experimental mean reward. Otherwise the scheduler selects channel i with probability p_i computed by (5).

In order to intuitively show the evolution of queue length under the channel scheduling policy designed by UE-IDS, an example of queue evolution is shown in Fig. 2.

In Fig. 2, the initial queue length was $Q(0) = 0$. At $t = 1$, the system entered the first busy period ($d = 1$), the channel with highest experimental mean reward was chosen. The following two slots scheduled channels based on IDS and the queue finally became empty at $t = 4$. The second idle period lasted for two slots ($t = 4 \sim 5$) and the second busy period lasted for seven slots ($t = 6 \sim 12$). In the third busy period, the queue was cleared within q_d slots, exploitation-only is enough to stabilize the system.

Algorithm 1 Uniformly Exploration and Information Directed Stabilizing Algorithm

Initialization: For each channel i, set $s_i(t), f_i(t), Q(t), d \leftarrow 0$.

for $t = 1, 2, \cdots T$ **do**
  Get $Q(t)$ by equation (1).
  if $Q(t) = 0$ **then**
    {idle period}
    Select channel $i \in K$ uniformly and randomly.
    Get feedback $X_i(t)$, $s_i(t) \leftarrow s_i(t-1) + X_i(t)$, $f_i(t) \leftarrow f_i(t-1) + 1 - X_i(t)$.
  else
    {busy period}
    if $Q(t-1) = 0$ **then**
      $d \leftarrow d + 1$, $l \leftarrow 0$
    end if
    $l \leftarrow l + 1$
    if $l \le q_d$ **then**
      Select channel $i = \arg\max_{i\in[K]} \frac{s_i(t)}{s_i(t)+f_i(t)}$.
    else
      Calculate $\Delta_i(t)$ and $g_i(t)$ by algorithm 2. Compute $p_i(t)$ by (5) and select channel i with probability p_i.
    end if
    Get feedback $X_i(t)$, $s_i(t) \leftarrow s_i(t-1) + X_i(t)$, $f_i(t) \leftarrow f_i(t-1) + 1 - X_i(t)$.
  end if
end for

Algorithm 2 Information Ratio Calculation$(s_i(t), f_i(t), K)$

$h_i(x) \leftarrow \text{Betapdf}(x, s_i(t), f_i(t))$
$H_i(x) \leftarrow \text{Betacdf}(x, s_i(t), f_i(t))$
$\overline{H}(x) \leftarrow \prod_i H_i(x)$
$G_i(x) \leftarrow \int_0^x y h_i(y) \mathrm{d}y$
$\Pr(i^* = i) \leftarrow \int_0^1 \frac{h_i(x)}{H_i(x)} \overline{H}(x) \mathrm{d}x$
$M_{i|i} \leftarrow \frac{1}{\Pr(i^*=i)} \int_0^1 \frac{x h_i(x)}{H_i(x)} \overline{H}(x) \mathrm{d}x$
$M_{i|i'} \leftarrow \frac{1}{\Pr(i^*=i')} \int_0^1 \frac{h_{i'}(x)\overline{H}(x)}{H_{i'}(x)H_i(x)} G_i(x) \mathrm{d}x, i' \neq i$
$\rho^* \leftarrow \sum_i \Pr(i^* = i) M_{i|i}$
$\Delta_i(t) \leftarrow \rho^* - \frac{s_i(t)}{s_i(t)+f_i(t)}$
$g_i(t) \leftarrow \sum_{i'} \Pr(i^* = i')(M_{i|i'} \log(M_{i|i'} \frac{s_i(t)+f_i(t)}{s_i(t)}) + (1 - M_{i|i'}) \log((1 - M_{i|i'}) \frac{s_i(t)+f_i(t)}{f_i(t)})$
return $\Delta_i(t), g_i(t)$.

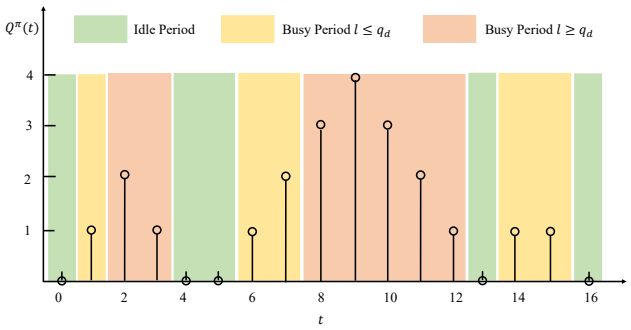

Fig. 2. An Example of Queue Evolution using UE-IDS

IV. Simulation Results

In this section, we provide some numerical simulations to validate the performance of the proposed algorithm. We consider a transmitter-receiver pair with $K = 4$ channels, the set of corresponding transmission success probabilities (service rates) is $\vec{\mu} = [0.3, 0.5, 0.7, 0.9]$. The algorithm is ran for $T = 3000$ consecutive slots. In Algorithm 2, $h_i(x)$ and $H_i(x)$ are continuous functions, it's hard to store continuous functions or compute the integrals of them. Therefore we compute the integral part for value $h_i(x)$, $H_i(x)$, $\overline{H}(x)$ and $G_i(x)$ through the Newton-Leibniz algorithm by dividing interval $[0, x]$ into $\tau = 2000$ intervals with uniform length.

We compare our algorithm with three other algorithms for solving the channel selection problem: (1). the classical stochastic bandit algorithm UCB1 [10] that chooses $i(t) := \arg\max_{i\in[K]}(\frac{s_i(t)}{s_i(t)+f_i(t)} + \sqrt{\frac{2\ln t}{s_i(t)+f_i(t)}})$ for each slot $t, t \ge K$; (2) the busy UCB1 algorithm that only runs UCB1 algorithm during busy periods, and does not send packets to test the channel during idle periods; (3) the UCB-UE algorithm proposed in [8] that balance explore-exploit trade-off when the queue length exceeds a certain threshold.

Fig. 3 examines the expected cumulative queue length regret of different algorithms. In this scenario, the arrival rate is set to $\lambda = 0.7$ and each plot shows the queue length regret over 1000 runs. It can be seen that our designed algorithm has much lower queue length regret than other algorithms, thanks to the fully explorations in idle periods and the properly use of IDS. Moreover, as time T goes up after $T = 1000$, the cumulative queue length grows at negligible speed, which validates that the proposed UE-IDS algorithm achieves a cumulative queue length regret of $\mathcal{O}(1)$ cumulative queue length regret.

In Fig. 4, we set the arrival rate as $\lambda = 0.95$, and the set of service rates is $\vec{\mu} = [0.3, 0.5, 0.7, 0.8]$, in such case, the arrival rate is larger than any of K service rates. The queue length regret is averaged over 100 simulation runs. The simulation result shows that the queue length regret of all algorithms keeps increasing, but our proposed UE-IDS has a much slower growing speed compared to three other algorithms. UE-IDS achieves smaller regret because it obtains a better explore-exploit trade-off when busy periods appear frequently under overloaded circumstance.

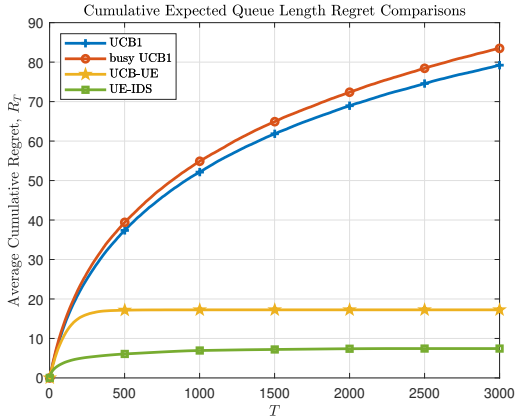

Fig. 3. Average Expected Queue Length Regret Comparisons of Different Algorithms over 1000 Trials with $\lambda = 0.7$

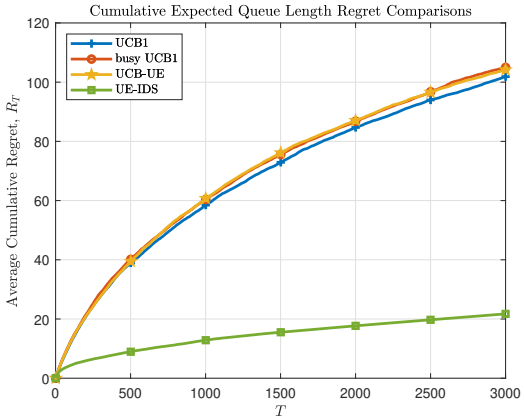

Fig. 4. Average Expected Queue Length Regret Comparisons of Different Algorithms over 100 Trials with $\lambda = 0.95$

V. Conclusions

In this paper, we study the online channel selection problem for a point-to-point communication link. To minimize the transmission delay in the absence of channel statistics, we propose an online channel selection algorithm that performs uniformly exploration during the idle periods, and achieve the explore-exploit trade-off using IDS during busy periods. Simulations show that our algorithm can achieve an optimal order $\mathcal{O}(1)$ regret when the packet arrival rate is within the stability region of the system. Future work includes extending the proposed method to block fading or fast fading channel conditions and completing theoretic analysis.

Appendix A
Details of Algorithm 1

In this appendix, we will introduce the computation procedures of the information gain, which requires the computation

of $\Delta_i(t)$ and $g_i(t)$. Recall that the posterior distribution of μ_i follows Beta distribution θ_i, we will use the posterior distribution to compute the probability of one channel having the highest reward and the conditional expected reward given the optimal channel, based on which $\Delta_i(t)$ and $g_i(t)$ can be computed.

First let $h_i(x) := \Pr(\theta_i = x|\mathcal{F}_{t-1})$ be the probability density function, $x \in [0,1]$, then the cumulative density function is denoted by $H_i(x) := \int_0^x h_i(t)dt = \Pr(\theta_i \leq x)$, which equals the probability that the mean reward of channel i is not more than x.

Let $\overline{H}(x) := \Pr(\theta_i \leq x, \forall i \in [K]) = \prod_i H_i(x)$ be the probability that the expected reward of all actions is not greater than x. The expectation of channel i has a mean reward not greater than x is denoted by $G_i(x) := \mathbb{E}[\theta_i \leq x] = \int_0^x y h_i(y)\mathrm{d}y$.

We can then compute the probability that channel i has the largest mean reward:

$$\begin{aligned}\Pr(i^* = i) &= \Pr(\theta_i = \max_j \theta_j) \\ &= \int_0^1 \Pr(\theta_i = x)\Pr(\theta_{i'} \leq x, \forall i' \neq i)\mathrm{d}x \\ &= \int_0^1 \frac{h_i(x)}{H_i(x)}\overline{H}(x)\mathrm{d}x.\end{aligned} \tag{6}$$

Let the expected mean reward of i given that i is the optimal choice be $M_{i|i}$, which can be computed by:

$$\begin{aligned}M_{i|i} &:= \mathbb{E}[\theta_i|i^* = i] \\ &= \int_0^1 x\frac{\Pr(\theta_i = x)\Pr(\theta_{i'} \leq x)}{\Pr(i^* = i)}\mathrm{d}x \\ &= \frac{1}{\Pr(i^* = i)}\int_0^1 \frac{x h_i(x)}{H_i(x)}\overline{H}(x)\mathrm{d}x.\end{aligned} \tag{7}$$

Similarly, denote the expected mean reward of i given that i' is the optimal choice by $M_{i|i'}$, which can be computed by:

$$\begin{aligned}M_{i|i'} &:= \mathbb{E}[\theta_i|i^* = i'] \\ &= \int_0^1\int_0^x \frac{y\Pr(\theta_i = y)\Pr(\theta_{i'} = x)\Pr(\theta_j \leq x, \forall j \neq i, i')}{\Pr(i^* = i')}\mathrm{d}y\mathrm{d}x \\ &= \frac{1}{\Pr(i^* = i')}\int_0^1 \frac{h_{i'}(x)\overline{H}(x)}{H_{i'}(x)H_i(x)}G_i(x)\mathrm{d}x, i' \neq i.\end{aligned} \tag{8}$$

Denote the expected mean reward of the optimal arm by ρ^*, which can be computed by:

$$\begin{aligned}\rho^* &:= \mathbb{E}[\max_i \theta_i] \\ &= \sum_i \Pr(i^* = i)\mathbb{E}[\theta_i|i^* = i] \\ &= \sum_i \Pr(i^* = i)M_{i|i}.\end{aligned} \tag{9}$$

As mentioned, $\Delta_i(t)$ measures the one-step difference between expected mean reward of the optimal arm and the experimental mean reward of the scheduled arm. The computation of $\Delta_i(t)$ is as follows:

$$\begin{aligned}\Delta_i(t) &:= \mathbb{E}[X_{i^*}(t) - X_i(t)|\mathcal{F}_{t-1}] \\ &= \rho^* - \hat{\mu_i} \\ &= \rho^* - \frac{s_i(t)}{s_i(t) + f_i(t)}.\end{aligned} \tag{10}$$

The information gain denoted by $g_i(t)$ is the mutual information between the optimal arm and the transmission outcome. In our system, to obtain $g_i(t)$, we need to calculate the expected Kullback-Leibler divergence of two Bernoulli distributions with parameter $M_{i|i'}$ and $\hat{\mu_i}$ as follows:

$$\begin{aligned}g_i(t) &:= I(i^*; X_i(t)) \\ &= \sum_{i' \in [K]} \Pr(i^* = i')D_{KL}(\Pr(X_i(t)|i^* = i')||\Pr(X_i(t))) \\ &= \sum_{i'} \Pr(i^* = i')(M_{i|i'}\log(M_{i|i'}\frac{s_i(t) + f_i(t)}{s_i(t)}) \\ &+ (1 - M_{i|i'})\log((1 - M_{i|i'})\frac{s_i(t) + f_i(t)}{f_i(t)}).\end{aligned} \tag{11}$$

Information gain measures the distance between the reward distributions with or without the knowledge of optimal arm. After computing all equations above, we can further compute the information ratio by equation (5).

REFERENCES

[1] H. Zhang, Y. Zhang, J. Cosmas, N. Jawad, W. Li, R. Muller, and T. Jiang, "mmwave indoor channel measurement campaign for 5g new radio indoor broadcasting," *IEEE Transactions on Broadcasting*, pp. 1–14, 2022.

[2] W. Aiello, E. Kushilevitz, R. Ostrovsky, and A. Rosén, "Adaptive packet routing for bursty adversarial traffic," *Journal of Computer and System Sciences*, vol. 60, no. 3, pp. 482–509, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022000099916811

[3] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," 1992.

[4] S. Krishnasamy, P. T. Akhil, A. Arapostathis, S. Shakkottai, and R. Sundaresan, "Augmenting max-weight with explicit learning for wireless scheduling with switching costs," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 2017, pp. 1–9.

[5] L. Huang, X. Liu, and X. Hao, "The power of online learning in stochastic network optimization," in *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 153–165. [Online]. Available: https://doi.org/10.1145/2591971.2591990

[6] P. Varaiya, J. Walrand, and C. Buyukkoc, "Extensions of the multiarmed bandit problem: The discounted case," *IEEE Transactions on Automatic Control*, vol. 30, no. 5, pp. 426–439, 1985.

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

[8] T. Stahlbuhk, B. Shrader, and E. Modiano, "Learning algorithms for minimizing queue length regret," *IEEE Transactions on Information Theory*, vol. 67, no. 3, pp. 1759–1781, 2021.

[9] Daniel, Russo, Benjamin, Van, and Roy, "Learning to optimize via information-directed sampling," *Operations Research: The Journal of the Operations Research Society of America*, 2018.

[10] P. Auer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, 2002.