Learning for Serving Deadline-Constrained Traffic in Multi-Channel Wireless Networks

Semih Cayci
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, Ohio 43210
Email: cayci.1@osu.edu

Atilla Eryilmaz
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, Ohio 43210
Email: eryilmaz.2@osu.edu

***Abstract*—We study the problem of serving randomly arriving and delay-sensitive traffic over a multi-channel communication system with time-varying channel states and unknown statistics. This problem deviates from the classical exploration-exploitation setting in that the design and analysis must accommodate the dynamics of packet availability and urgency as well as the cost of each channel use at the time of decision. To that end, we have developed and investigated two policies, one index-based (UCB-Deadline) and the other Bayesian (TS-Deadline), both of which perform dynamic channel allocation decisions that incorporate these traffic requirements and costs. Under symmetric channel conditions, we have proved that the UCB-Deadline policy can achieve bounded regret in the likely case where the cost of using a channel is not too high to prevent all transmissions, and logarithmic regret otherwise. In our numerical studies, we also show that TS-Deadline achieves superior performance over its UCB counterpart, making it a potentially useful alternative when fast convergence to optimal is important.**

I. Introduction

With the advances in wireless communications, next generation communication networks are expected to serve real-time applications that require end-to-end deadline constraints and a large amount of throughput over fading channels. Especially real-time multimedia applications such as voice and video streaming possess stringent deadline constraints that require particular emphasis. The ultra-wideband communication channels that are designed to meet these requirements, such as millimeter-wave (mmW) channels, have highly intermittent dynamics, which makes existing channel probing and estimation techniques inapplicable. Therefore, it is crucial to develop new communication schemes that can handle applications with deadline constraints and large throughput demands in the absence of channel statistics and channel state information.

In wireless communication schemes such as IEEE 802.11 and 5G millimeter-wave (mmW) cellular systems, availability of multiple orthogonal channels enables a user to simultaneously utilize multiple channels to increase the quality of communication in various aspects [1], [2]. In [1], it is shown that multi-channel operation provides significant increase in network capacity, which can be exploited to meet the increasing demand for throughput. In mmW cellular communications, multi-channel scenario is expected to overcome the intermittence problem of mmW channels due to blockage, which particularly hinders applications with quality of service (QoS) requirements [2], [3]. As it is possible to equip a single node with multiple radio interfaces due to the reduced hardware costs, multi-channel communication scheme offers a feasible solution to serve applications with deadline constraints and large throughput demand [1], [4]. On the other hand, operational costs, such as power consumption, impose a critical constraint in the number of active interfaces. Thus, it is important to activate a plausible number of channels dynamically depending on queue-length and deadline constraints so as to increase throughput while keeping the operational costs at acceptable levels.

In conventional communication systems, there are efficient channel estimation techniques that provide channel state information (CSI) for rate and power allocation policies [5]. However, these methods are inapplicable in millimeter-wave communication systems as the channels are highly intermittent and fast-varying [2], [6], [7], [8]. This necessitates the development of online learning algorithms that rely on channel feedback in the absence of channel state information and channel statistics.

In this paper, we investigate the problem of dynamic channel allocation for a single user in a multi-channel network with deadline constraints and service costs in the absence of channel statistics and CSI. Our main contribution is two online learning algorithms that converge to the optimal solutions with small regret by using only the channel feedback. In traditional communication systems, efficient rate and power allocation schemes that base the decisions on CSI and queue-lengths exist [9], [10], [11], [12], [13], [14]. However, these methods are built on the key assumption that CSI is available at the time of decision, therefore they are not applicable in the emerging communication scenarios where CSI and channel statistics are unknown. There is an interesting body of work which considers the online learning problem for rate allocation based on success/fail feedback [15], [16]. These works do not apply to our context since they do not provide short-term performance guarantees, such as regret.

This work is funded by the NSF grants: CCSS-EARS-1444026, CNS-NeTS-1514127, CMMI-SMOR-1562065 and CNS-WiFiUS-1456806, CNS-ICN-WEN-1719371; the DTRA grant HDTRA1-15-1-0003; and the QNRF Grant NPRP 7-923-2-344.

978-3-901882-90-6/17 ©2017 IFIP

There is a large body of work in the design and analysis of online learning algorithms that optimize short-term performance in the context of multi-armed bandits (MAB) [17], [18]. Our work deviates from the context of classical stochastic bandits as the revenue of the activated arms are coupled, the controller has the incentive to activate no channels due to the cost, and there is a strong dependence on the queue-length. In [20], learning problem is investigated with a regret definition based on queueing-delay. This work does not apply to our setting as it does not consider deadline-constrained traffic.

The organization of this paper is as follows. In Section II, we present the system model for the multi-channel network with deadline constraints and service costs along with the definitions of admissibility, optimal policy and regret. In Section III, first we characterize the optimal policy for a particular definition of throughput, then we propose two online learning algorithms, one index-based (UCB-Deadline) and the other Bayesian (TS-Deadline), that achieve bounded regret in the interesting case that cost is small enough for channel use. In Section IV, we provide performance guarantees for UCB-Deadline. Finally, in Section V, we evaluate the regret performances of the proposed algorithms in specific communication scenarios.

II. System Model

We consider a discrete-time system. The packets arrive into the system according to an arrival process $A(t)$ which is independent and identically distributed (iid) over a finite set $\mathcal{A} = \{0, 1, \ldots, A_{max}\}$ with probability distribution $\mathbb{P}(A(t) = a) = \alpha_a$ for $a \in \mathcal{A}$. The packets have a lifetime of one timeslot, and will be lost if they are not served immediately.

The packets in the queue can be transmitted by K (possibly infinite) independent fading channels. The rate $C_k(t)$ of channel k evolves according to an iid Bernoulli process with mean μ, i.e., $C_k(t) \overset{iid}{\sim} Ber(\mu)$ for $k = 1, 2, \ldots, K$. This Bernoulli channel model reflects the sharp difference between line-of-sight (LOS) and non-line-of-sight channel (NLOS) states in millimeter-wave communications [2], [6], [7], [8]. $C_k(t)$ is revealed via ACK or NACK signals after the transmission only if channel k is activated at time t. The system is illustrated in Figure 1.

In order to increase the reliability of communication under channel uncertainty, more channels than the queue-length can be activated at every timeslot. The number of channels to be activated at a timeslot is determined by a centralized controller who has the knowledge of queue-length prior to the decision and also that all channels are Bernoulli distributed with the same mean, μ, which is unknown and learned over time. Each channel use incurs a constant cost of $d \in [0, 1]$, which measures the operational costs associated with each channel use, such as power. It is assumed that d is known by the controller. If k channels are activated and there are a packets in the queue at timeslot t, the revenue is as follows:

$$X_{(a,k)}(t) = \tau_{(a,k)}(t) - k \cdot d, \tag{1}$$

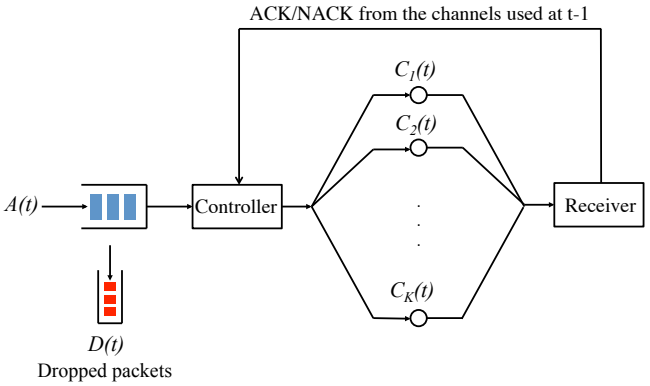

Fig. 1. The multi-channel network with symmetric Bernoulli channels. At the end of each transmission, CSI of each activated channel is revealed to the controller via ACK/NACK signals.

where $\tau_{(a,k)}(t)$ is the throughput at time t when a packets arrive and k channels are activated. Note that $\tau_{(a,k)}(t) \leq a$ for any k and a, which implies that $X_{(a,k)}(t) < 0$ for all $a \in \mathcal{A}$ and $k > \frac{A_{max}}{d}$ since throughput is bounded and the cost of adding a new channel increases linearly. The maximum number of channels that can be activated at any timeslot is denoted by $K_{max} = \min\{K, \frac{A_{max}}{d}\}$. An example for τ is the number of successfully transmitted packets out of a, which has the following expectation:

$$\mathbb{E}\tau_{(a,k)}(t) = \begin{cases} k\mu, & \text{if } k \leq a \\ \sum_{i=1}^{k} (i \wedge a)\binom{k}{i}(1-\mu)^{k-i}\mu^i, & \text{if } k > a. \end{cases} \tag{2}$$

The decision variable is the number of activated channels at each timeslot. Let $I_a(t)$ be the number of activated channels at time t when there are $a \in \mathcal{A}$ packets in the queue. Then, the policy I is defined as follows:

$$I_{A(t)}(t) = \sum_{a \in \mathcal{A}} \mathbb{I}_{\{A(t)=a\}} I_a(t), \tag{3}$$

where $\mathbb{I}$ is the indicator function. Also, let $\mathcal{U}(t)$ denote the set of channels activated at time t. An admissible policy I is based on the knowledge of channel realizations until and excluding t, and arrivals until and including t:

$$\mathbb{I}_{\{I_{A(t)}(t)=k\}} \in \sigma(\{C_i(s) : i \in \mathcal{U}(s), s < t\}, \{A(s) : s \leq t\}),$$

for all $k = 1, 2, \ldots, K$, where $\sigma(\{X_i\}_{i=1}^{J})$ denotes the σ-field generated by a collection of random variables X_i, $i = 1, 2, \ldots, J$.

If a genie reveals the mean μ to the controller, the optimal policy that maximizes the revenue given that $A(t) = a$ is as follows:

$$I_a^*(\mu) = \max_{1 \leq k \leq K} \mathbb{E}[X_{(a,k)}(t)]. \tag{4}$$

Note that for fixed a and k, $\mathbb{E}[X_{(a,k)}(t)]$ is constant over time because of the iid nature of fading channels. Therefore, the time index will be dropped in such cases.

As the a priori knowledge of μ is absent, an algorithm has to learn the mean, and maximize the revenue simultaneously. Pseudo-regret, which will be simply referred to as regret throughout this paper, is a common measure to evaluate the performance of learning algorithms [17], [19], [18]. The regret under an admissible policy I is defined as follows:

$$\begin{aligned}\bar{R}_n = \ & \mathbb{E}\Big[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}}\mathbb{I}_{\{A(t)=a\}}(X_{(a,I^*_a(\mu))}(t)-X_{(a,I_a(t))}(t))\Big] \\ = \ & \sum_{a\in\mathcal{A}}\alpha_a\cdot\mathbb{E}\Big[\sum_{t=1}^{n}(X_{(a,I^*_a(\mu))}(t)-X_{(a,I_a(t))}(t))\Big].\end{aligned} \tag{5}$$

In words, regret is defined as the cumulative difference between the maximum expected revenue given the mean μ and the expected revenue under policy I in n timeslots.

The objective in this paper is to design policies that provide low regret. In the following section, we propose algorithms that achieve this goal.

III. Optimal Policy and Algorithm Design

In this section, we will investigate the behavior of the optimal policy and introduce two algorithms that achieve desirable regret performances for the exploration-and-exploitation problem in the deadline-constrained multi-channel network described in the previous section.

A. Optimal Policy

The optimal policy is illustrated as a function of μ in Figure 2 under the throughput given in (2) for the case $A_{max} = 6$ and $d = 0.2$.

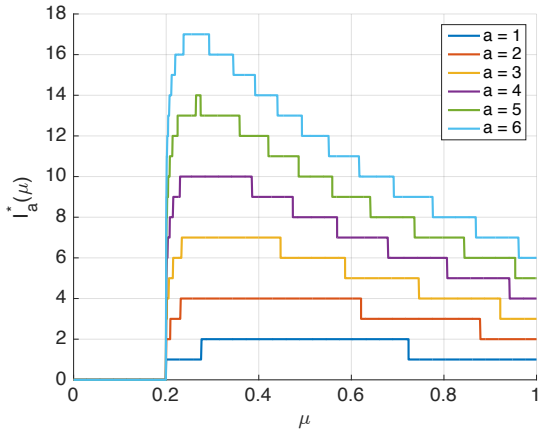

Fig. 2. Optimal policy as a function of μ for the case $A_{max} = 6$ and $d = 0.2$. The behavior is highly dependent on the queue-length a.

Note that the optimal number of channels to maximize revenue increases up to some μ to increase reliability, then decreases since an additional channel is too costly as the current ones are reliable enough. Also, note that $I^*_0(\mu) = 0$ for all $\mu \in [0, 1]$.

B. UCB-Deadline

For the exploration-and-exploration problem at hand, learning must be reinforced when the confidence is low to avoid linear regret in certain sample paths on which exploration is stopped at an early stage, and the estimates must converge to the true mean after a sufficiently long time for achieving small regret in the long-run. Utilization of upper confidence bound (UCB) in the absence of the true mean reinforces learning through "optimism in the face of uncertainty" [17], therefore is a suitable strategy in algorithm design. In the following, we define a policy named UCB-Deadline that makes use of UCB to determine the number of channels to be activated.

Definition 1 (UCB-Deadline)**.** Let $T(t) = \sum_{s=1}^{t} I_{A(s)}(s)$ be the number of activated channels until timeslot t,

$$m(t) = \frac{1}{T(t)}\sum_{s=1}^{t}\sum_{i\in\mathcal{U}(s)} C_i(s), \tag{6}$$

be the sample mean of activated channels until timeslot t, and $c_{t,s} = \sqrt{\frac{\beta\log t}{2s}}$ for $\beta > 0$. UCB at timeslot t is defined as follows:

$$\bar{\mu}_{T(t-1)}(t) = m(t-1) + c_{t,T(t-1)}. \tag{7}$$

Let $\hat{\mu}_s$ be the sample mean of channel realizations after s channel uses. Since all channels are iid and symmetric, $m(t-1) \overset{d}{=} \hat{\mu}_{T(t-1)}$, which will provide simplicity in the performance analysis.

With these definitions, UCB-Deadline with parameter β, denoted as UCB-Deadline(β), is summarized in Algorithm 1,

Algorithm 1: UCB-Deadline(β)

input: $\beta > 0$

Initialization: $T(0) = 1$; $m(0) \sim Ber(\mu)$;
for $t = 1, 2, \ldots, n$ **do**
$\bar{\mu}_{T(t-1)}(t) = m(t-1) + c_{t,T(t-1)}$;
$I_{A(t)}(t) = I^*_{A(t)}(\bar{\mu}_{T(t-1)}(t))$;
$T(t) = T(t-1) + I_{A(t)}(t)$;
$m(t) = \frac{1}{T(t)}\cdot\Big(T(t-1)\cdot m(t-1) + \sum_{i=1}^{I_{A(t)}(t)} C_i(t)\Big)$;

where $(I^*_a)_{a\in\mathcal{A}}$ is the optimal policy defined in (4).

C. TS-Deadline

In problems that involve exploration-and-exploitation trade-off, Thompson Sampling provides effective solutions that reinforce learning through randomization [21], [22]. In the following, we propose an algorithmic prescription to the learning problem at hand based on Thompson Sampling, which is abbreviated as TS-Deadline.

Definition 2 (TS-Deadline)**.** Let $Beta(\theta_0, \theta_1)$ denote the beta distribution with parameters $\theta_i > 0$ for $i = 0, 1$ whose probability density function is given by $f(x; \theta_0, \theta_1) =$

Algorithm 2: TS-Deadline

Initialization: $\theta_0(0) = 1,\ \theta_1(0) = 1$
for $t = 1, 2, \ldots, n$ **do**
$\bar{\mu}^{TS}(t) \sim Beta(\theta_0(t-1), \theta_1(t-1))$;
$I_{A(t)}(t) = I^*_{A(t)}(\bar{\mu}^{TS}(t))$;
$\theta_k(t) = \theta_k(t-1) + \sum_{i=1}^{I_{A(t)}(t)} \mathbb{I}_{\{C_i(t)=k\}},\ k = 0, 1.$

$\frac{\Gamma(\theta_0+\theta_1)}{\Gamma(\theta_0)\Gamma(\theta_1)}(1-x)^{\theta_0-1}x^{\theta_1-1}$ [21]. TS-Deadline is described in Algorithm 2.

In the following section, performance guarantees under UCB-Deadline will be presented in the form of regret upper bounds.

IV. Performance Analysis of UCB-Deadline

In this section, we will provide upper bounds for the regret under UCB-Deadline. The strategy to accomplish this is as follows: first we will provide two lemmas in a general setting, and then use these lemmas to upper bound the regret under UCB-Deadline.

Lemma 1. *Consider a case where the optimal policy is $I^*_a(\mu) = k_a \cdot \mathbb{I}_{\{\mu > d\}}$ for some $k_a \in \{1, 2, \ldots, K_{max}\},\ \forall a \in \mathcal{A}$. Let $T_0(n) = \sum_{t=1}^{n} \mathbb{I}_{\{I_{A(t)}(t)=0\}}$ be the number of timeslots when all channels are idle under UCB-Deadline. Under UCB-Deadline with $\beta \geq 3$, the following upper bounds are obtained for any $a \in \mathcal{A}$:*

1) *If $\mu > d$, then $\mathbb{E}[T_0(n)] \leq K_{max}\frac{\pi^2}{6}$,*
2) *If $\mu \leq d$, then*

$$\mathbb{E}[n - T_0(n)] \leq \frac{2\beta \log n}{(d-\mu)^2} + K_{max}\frac{\pi^2}{6},$$

for all $n \geq 1$.

Lemma 1 implies that in a binary decision case, UCB-Deadline makes a bounded number of wrong decisions if the true mean is higher than the cost, and a logarithmically growing number of wrong decisions over time otherwise in the expected sense.

Lemma 2. *Fix $a \in \mathcal{A}$. Let $\tau^l_a < \tau^u_a$ be two given constants in $[0, 1]$. Consider the following case:*

$$I^*_a(\mu) = \begin{cases} 0, & \text{if } \mu < \tau^l_a \\ k^*_a, & \text{if } \mu \in [\tau^l_a, \tau^u_a] \\ k_a, & \text{if } \mu > \tau^u_a. \end{cases} \tag{8}$$

for some $k_a, k^*_a > 0$. Assume $\mu \in [\tau^l_a, \tau^u_a]$. Under UCB-Deadline with $\beta \geq 4$, the following upper bounds hold for all $n \geq 1$:

1) $\mathbb{E}[\sum_{t=1}^{n} \mathbb{I}_{\{I_a(t)=0\}}] \leq K_{max}\frac{\pi^2}{6}$.
2) $\mathbb{E}[\sum_{t=1}^{n} \mathbb{I}_{\{I_a(t)=k_a\}}] \leq K_{max}\frac{\pi^2}{6} + M(\frac{(\mu-\tau^u_a)^2}{2\beta}) < \infty$, *where*

$$M(\epsilon) = t_\epsilon + K_{max} \cdot \frac{\pi^2}{2} \sum_{t=1}^{\infty} \frac{1}{(t - \log(t+1)/\epsilon)^2}$$

and $t_\epsilon = \inf\{t : t - \frac{\log(t+1)}{\epsilon} > 0\}$.

Lemma 2 says that if the true mean is in an interval with nonempty interior so that the correct decision can be made after sufficient concentration around the mean, then the numbers of wrong decisions under UCB-Deadline are bounded in both directions in the expected sense.

Proofs of Lemma 1 and Lemma 2 will be given in Appendix.

The following theorem provides performance guarantees under UCB-Deadline.

Theorem 1 (Regret Upper Bounds for UCB-Deadline)**.** *Let $k^{min}_a = \underset{k=0,1,\ldots,K_{max}}{\arg\min}\ \mathbb{E}X_{(a,k)},\ \forall a \in \mathcal{A}$. Then, the following upper bounds hold for the regret under UCB-Deadline with parameter $\beta \geq 4$.*

1) *If $\mu < d$, then*

$$\bar{R}_n \leq \sum_{a \in \mathcal{A}} \alpha_a(-\mathbb{E}X_{(a,k^{min}_a)}) \cdot \Big(\frac{2\beta \log n}{(d-\mu)^2} + K_{max}\frac{\pi^2}{6}\Big). \tag{9}$$

2) *If $\mu > d$, let $\tau^l_a \leq \mu \leq \tau^u_a$ be the largest interval such that $I^*_a(\tilde{\mu}) = I^*_a(\mu),\ \forall \tilde{\mu} \in [\tau^l_a, \tau^u_a]$ for any $a \in \mathcal{A}$. Then,*

$$\bar{R}_n \leq \sum_{a \in \mathcal{A}} \alpha_a \Big(\mathbb{E}X_{(a,I^*_a(\mu))} - \mathbb{E}X_{(a,k^{min}_a)}\Big) \cdot \Big(K_{max}\frac{\pi^2}{3} + M\Big(\frac{(\mu-\tau^u_a)^2}{2\beta}\Big)\Big). \tag{10}$$

Proof. 1) Note that $\mu < d$ implies $I^*_a(\mu) = 0,\ \forall a \in \mathcal{A}$. Thus, the regret is upper bounded by using (5) as follows:

$$\begin{aligned}
\bar{R}_n &= \sum_{a \in \mathcal{A}} \alpha_a \cdot \mathbb{E}\Big[\sum_{t=1}^{n} -X_{(a,I_a(t))}(t)\Big] \\
&\leq \sum_{a \in \mathcal{A}} \alpha_a \cdot \mathbb{E}[\sum_{t=1}^{n} -X_{(a,k^{min}_a)}(t) \cdot \mathbb{I}_{\{I_a(t)\neq 0\}}] \\
&\overset{(a)}{=} \sum_{a \in \mathcal{A}} \alpha_a \mathbb{E}[-X_{(a,k^{min}_a)}] \cdot \mathbb{E}[\sum_{t=1}^{n} \mathbb{I}_{\{I_a(t)\neq 0\}}] \\
&= \sum_{a \in \mathcal{A}} \alpha_a \mathbb{E}[-X_{(a,k^{min}_a)}] \cdot \mathbb{E}[n - T_0(n)] \\
&\overset{(b)}{\leq} \sum_{a \in \mathcal{A}} \alpha_a \mathbb{E}[-X_{(a,k^{min}_a)}] \cdot (\frac{2\beta \log n}{(d-\mu)^2} + K_{max}\frac{\pi^2}{6}),
\end{aligned}$$

where (a) follows from the fact that $X_{(a,k^{min}_a)}(t)$ is a function of $\{C_k(t)\}_{k=1}^{K_{max}}$ and therefore independent from $I_a(t)$ and its expectation is time-invariant, and (b) follows from Lemma 1.

2) Let $\Delta X^{max}_a = \mathbb{E}[X_{(a,I^*_a(\mu))}(t) - X_{(a,k^{min}_a)}(t)]$ be the maximum expected error for any $a \in \mathcal{A}$ and $k \in$

$\{0, 1, \ldots, K_{max}\}$ at any timeslot t. Then, the regret under UCB-Deadline can be upper bounded as follows:

$$\begin{aligned}
\bar{R}_n &\leq \sum_{a\in\mathcal{A}} \alpha_a \Delta X_a^{max} \sum_{t=1}^{n} \mathbb{E}\Big[\mathbb{I}_{\{I_a(t)\neq I_a^*(\mu)\}}\Big] \\
&\leq \sum_{a\in\mathcal{A}} \alpha_a \Delta X_a^{max} \sum_{t=1}^{n} \mathbb{E}\Big[\mathbb{I}_{\{\bar{\mu}_{T(t-1)}(t)<\tau_a^l\}} \\
&\qquad + \mathbb{I}_{\{\bar{\mu}_{T(t-1)}(t)>\tau_a^u\}}\Big] \\
&\overset{(a)}{\leq} \sum_{a\in\mathcal{A}} \alpha_a \Delta X_a^{max} \sum_{t=1}^{n} \mathbb{E}\Big[\mathbb{I}_{\{I_a(t)=0\}} \\
&\qquad + \mathbb{I}_{\{I_a(t)=\min\limits_{\hat{\mu}>\tau_a^u} I_a^*(\hat{\mu})\}}\Big] \\
&\overset{(b)}{\leq} \sum_{a\in\mathcal{A}} \alpha_a \Delta X_a^{max} (K_{max}\frac{\pi^2}{3} + M(\frac{(\tau_a^u-\mu)^2}{2\beta})),
\end{aligned}$$

where (a) follows from the fact that minimal learning and maximal possible regret per timeslot maximize the overall regret, and (b) is a direct application of Lemma 2.

□

Theorem 1 implies that the regret under UCB-Deadline is bounded if channel usage is feasible and decision errors can be eliminated as a result of the concentration around the true mean after sufficiently many trials. This is an interesting result since the regret is logarithmic in most classical MAB settings whereas we have a bounded regret in this case [17], [18].

As an illustrative example, we will consider the case $A_{max} = 2$ and $K_{max} = 2$ in the rest of this section. For this case, the decision regions are given in Figure 3.

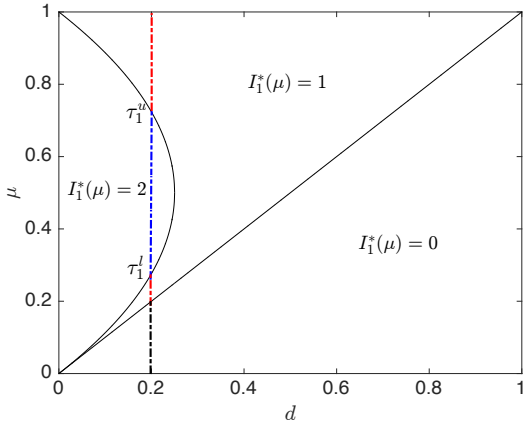

Fig. 3. Optimal decision regions $I_1^*(\mu)$ for the case of $A_{max} = 2$ and $K_{max} = 2$. The black, red and blue dashed lines indicate $I_1^*(\mu) = 0$, $I_1^*(\mu) = 1$, $I_1^*(\mu) = 2$, respectively for $d = 0.2$.

For this example, $\tau_1^l = \frac{1}{2} - \sqrt{\frac{1}{4} - d}$, $\tau_1^u = \frac{1}{2} + \sqrt{\frac{1}{4} - d}$, $\tau_2^l = d$ and $\tau_2^u = 1$. Note that the use of extra channel when $a = 1$ is infeasible for any μ when $d > 0.25$ in this case. For $d = 0.2$, corresponding decision regions $I_1^*(\mu)$ are illustrated in Figure 4.

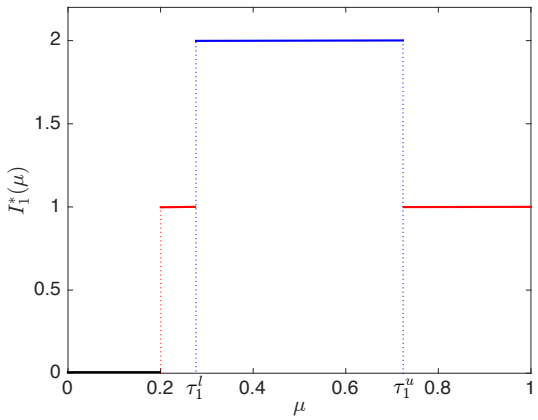

Fig. 4. Optimal decisions $I_1^*(\mu)$ for $d = 0.2$ and $K_{max} = 2$. Using an additional channel becomes infeasible when the channel is too reliable.

The optimal number of channels when $a = 1$ is $I_1^*(\mu) = 2$ when $\mu \in [\tau_1^l, \tau_1^u]$, and it decreases to 1 when $\mu > \tau_1^u$. This is observed because the additional channel becomes costly when the channel is highly reliable, i.e., μ is too high.

V. Numerical Results

In this section, we will provide simulation results for regrets under UCB-Deadline and TS-Deadline in a variety of scenarios. The revenue function in these simulations is chosen as the one defined in (2).

For $\mu = 0.52$ and $d = 0.2$, we first investigate the performance of a naïve pure exploitation algorithm called PE-Deadline, which makes decisions according to the sample means solely. The regret under PE-Deadline is given in Figure 5.

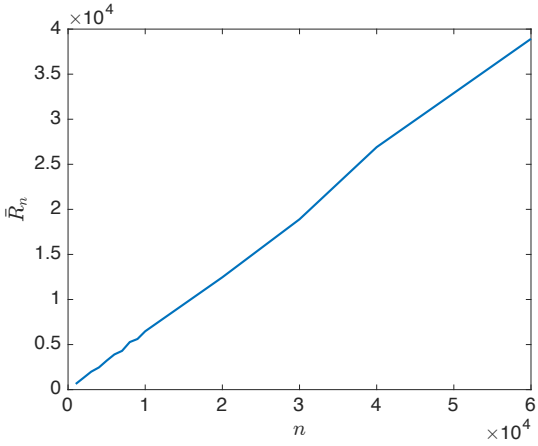

Fig. 5. Regret under PE-Deadline for the case $A_{max} = 6$, $d = 0.2$ and $\mu = 0.52$. Exploration under PE-Deadline halts in early stages with a non-zero probability, which leads to this linear regret behavior.

The regret under PE-Deadline grows linearly over time. This is because in the early stages of learning when confidence is low, PE-Deadline is prone to stop learning. This observation

emphasizes the necessity of reinforcement in learning, which is captured in UCB-Deadline and TS-Deadline.

The arrival process throughout this section is chosen as an iid process with the truncated Poisson distribution, denoted by $Poisson_{A_{max}}(\lambda)$ for $\lambda > 0$, which has the following probability mass function:

$$\mathbb{P}(A(t)=k) = \begin{cases} \frac{\lambda^k/k!}{\sum_{i=0}^{A_{max}} \lambda^i/i!}, & \text{if } 0 \le k \le A_{max} \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

where A_{max} is the maximum number of arrivals in a timeslot.

For $\mu = 0.52$ and $d = 0.2$, simulation results under UCB-Deadline and TS-Deadline are provided in Figure 6. The arrival distribution $\{\alpha_a\}_{a\in\mathcal{A}}$ is chosen as a truncated Poisson distribution with maximum element $A_{max} = 6$.

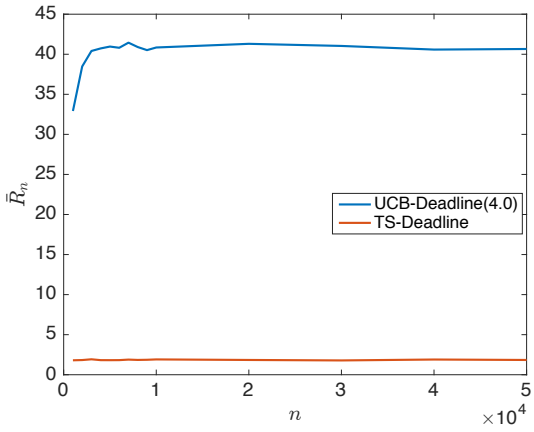

Fig. 6. Regrets under UCB-Deadline with parameter $\beta = 4$ and TS-Deadline for the case $A_{max} = 6$, $d = 0.2$ and $\mu = 0.52$. Since $\mu > d$, bounded regret is observed in both cases as expected.

From Figure 2, it is observed that for all $a \in \mathcal{A}$, there exists τ_a^l, τ_a^u such that $\mu \in (\tau_a^l, \tau_a^u)$ and the optimal decision is constant in that interval. Therefore, the regret is bounded by Theorem 1, which is verified by Figure 6. Also, it is noteworthy that TS-Deadline achieves significantly smaller regret than UCB-Deadline in this case.

Performance results for $\mu = 0.15$ and $d = 0.2$ are illustrated in Figure 7 with the same truncated Poisson distribution for the arrival process. Note that $\mu < d$ in this case, and therefore channel usage is infeasible for any queue-length. By Theorem 1, the upper bound for the regret under UCB-Deadline is logarithmic over time, consistent with the simulation results. It is observed that TS-Deadline also has an increasing regret, but it achieves significantly lower regret than UCB-Deadline in this case as well.

In order to observe the effect of queue-length on regret, we consider the case where $\mu = 0.52$, $d = 0.2$ are fixed and $A(t) \sim Poisson_{A_{max}}(3)$ for various values of A_{max}. Regret performances of UCB-Deadline with parameter $\beta = 4$ and TS-Deadline are given in Figure 8.

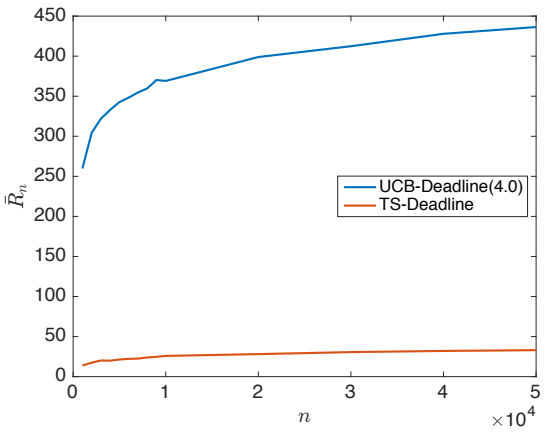

Fig. 7. Regrets under UCB-Deadline with parameter $\beta = 4$ and TS-Deadline for the case $A_{max} = 6$, $d = 0.2$ and $\mu = 0.15$.

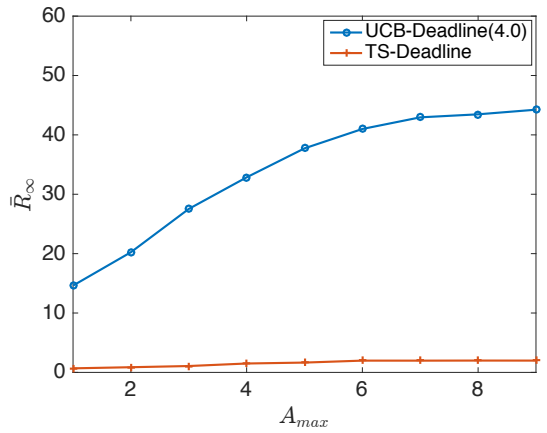

Fig. 8. $\bar{R}_\infty$ for various values of A_{max} when $A(t) \sim Poisson_{A_{max}}(3)$. The regret has an increasing behavior with a decreasing rate of increase with respect to A_{max} for fixed $\lambda = 3$.

For fixed λ, the regrets increase with the maximum queue-length A_{max} under both algorithms. However, the rate of increase has a decreasing behavior with respect to A_{max}.

VI. Conclusion

In this paper, we investigated the channel allocation problem in a wireless network under a deadline-constrained traffic when the channel statistics and channel state information are unknown. We first identified the optimal policy assuming that the channel statistics are known by the controller. Then, we proposed two learning algorithms, an index-based policy (UCB-Deadline) and a Bayesian policy (TS-Deadline). We proved that the regret under UCB-Deadline is bounded in the likely case that channel use is feasible, and logarithmic otherwise. This is an interesting result as the regret is logarithmic in most MAB problems. Our numerical investigations revealed that TS-Deadline achieves significantly lower regret

than UCB-Deadline, which suggests it is potentially a useful alternative for faster learning.

It is assumed that there is a single class of independent and statistically symmetric channels in this work. UCB-Deadline is proved to achieve a bounded regret by incorporating the number of pending packets and utilizing the knowledge of statistical symmetry of the channels. In an extension of this setting where there are multiple classes of statistically symmetric channels, a similar exploitation of statistical symmetry may provide significant performance improvements. As a future work, we would like to investigate the learning problem in this extended setting.

In this paper, we investigated the performance of TS-Deadline numerically and observed that it achieves significantly lower regret than its UCB counterpart. Another future work might be the regret analysis of TS-Deadline and comparison with UCB-Deadline.

On the side of the service, an interesting extension of this work might be the learning problem where certain QoS requirements such as delivery ratio and service regularity must be met.

APPENDIX

A. Proof of Lemma 1

1) If $\mu \geq d$,

$$\begin{aligned} T_0(n) &= \sum_{t=1}^{n} \mathbb{I}_{\{\bar{\mu}_{T(t-1)}(t)<d\}} \\ &\leq \sum_{t=1}^{n} \mathbb{I}_{\{\min\limits_{1\leq s\leq K_{max}\cdot t} \bar{\mu}_s(t)<d\}} \\ &\leq \sum_{t=1}^{n}\sum_{s=1}^{K_{max}\cdot t} \mathbb{I}_{\{\bar{\mu}_s(t)<d\}} \\ &\leq \sum_{t=1}^{n}\sum_{s=1}^{K_{max}\cdot t} \mathbb{I}_{\{\bar{\mu}_s(t)<\mu\}}. \end{aligned}$$

by following a similar path as [17]. Taking the expectation and using Hoeffding-Chernoff Bound, the following is obtained if $\beta \geq 3$:

$$\begin{aligned} \mathbb{E}[T_0(n)] &\leq K_{max}\sum_{t=1}^{n} t^{1-\beta} \\ &\leq K_{max}\sum_{t=1}^{\infty} t^{1-\beta} \leq K_{max}\frac{\pi^2}{6}. \end{aligned}$$

2) The following claim is necessary for proving this part.

Claim 1. *If $I_{A(t)}(t) > 0$, then at least one of the following must hold:*

a) $\hat{\mu}_{T(t-1)} \geq \mu + c_{t,T(t-1)}$
b) $T(t-1) < \frac{2\beta \log n}{(d-\mu)^2}$.

Proof of Claim 1. Suppose neither holds. Then,

$$\begin{aligned} \hat{\mu}_{T(t-1)} + c_{t,T(t-1)} &< \mu + 2c_{t,T(t-1)} \\ &\leq \mu + 2c_{n,T(t-1)} \\ &\leq \mu + (d-\mu) = d. \end{aligned}$$

Thus, $I_{A(t)}(t) = 0$. □

Let $l = \left\lceil \frac{2\beta \log n}{(d-\mu)^2} \right\rceil$. Then, by using a similar methodology as [17],

$$\begin{aligned} n - T_0(n) &\leq l + \sum_{t=1}^{n} \mathbb{I}_{\{\bar{\mu}_{T(t-1)}(t)>d, T(t-1)\geq l\}} \\ &= l + \sum_{t=1}^{n} \mathbb{I}_{\{\hat{\mu}_{T(t-1)}\geq \mu + c_{t,T(t-1)}\}} \\ &\leq l + \sum_{t=1}^{n}\sum_{s=1}^{K_{max}t} \mathbb{I}_{\{\hat{\mu}_{T(t-1)}\geq \mu + c_{t,T(t-1)}\}}, \end{aligned}$$

where the first line holds with equality iff $T(t-1) \geq l$ and the second line follows from Claim 1. Taking the expectation and exploiting Hoeffding-Chernoff Bound, the result is obtained.

B. Proof of Lemma 2

1) Proof of this part is similar to the proof of the first part of Lemma 1.
2) The following claims play an essential role in the proof.

Claim 2. *If $I_a(t) = k_a$, then at least one of the following must hold:*

a) $\hat{\mu}_{T(t-1)} > \mu + c_{t,T(t-1)}$,
b) $c_{t,T(t-1)} > \frac{\tau_a^u - \mu}{2}$.

Claim 3. *For any $n \geq 1$, UCB-P with parameter $\beta \geq 4$ provides the following:*

$$\mathbb{E}[T_0^2(n)] \leq K_{max}\frac{\pi^2}{2}.$$

Claim 4. *For any $\epsilon > 0$, UCB-P with $\beta \geq 4$ implies the following:*

$$\sum_{t=0}^{\infty} \mathbb{P}\Big(\frac{\log(t+1)}{T(t)} > \epsilon\Big) \leq M(\epsilon) < \infty$$

The proofs for Claim 2, Claim 3 and Claim 4 are given at the end of this subsection.

Let $\tilde{T}(n) = \sum_{t=1}^{n} \mathbb{I}_{\{I_a(t)=k_a\}}$. Using Claim 2, $\tilde{T}(n)$ can be upper bounded as follows:

$$\begin{aligned} \tilde{T}(n) &\leq \sum_{t=1}^{n} \mathbb{I}_{\{c_{t,T(t-1)}>\frac{\tau-\mu}{2} \text{ or } \hat{\mu}_{T(t-1)}>\mu+c_{t,T(t-1)}\}} \\ &\leq \sum_{t=1}^{n} \mathbb{I}_{\{c_{t,T(t-1)}>\frac{\tau-\mu}{2}\}} + \sum_{t=1}^{n} \mathbb{I}_{\{\hat{\mu}_{T(t-1)}>\mu+c_{t,T(t-1)}\}} \\ &= \sum_{t=1}^{n} \mathbb{I}_{\{\frac{\log t}{T(t-1)}>\frac{(\tau-\mu)^2}{2\beta}\}} + \sum_{t=1}^{n} \mathbb{I}_{\{\hat{\mu}_{T(t-1)}>\mu+c_{t,T(t-1)}\}}. \end{aligned}$$

Thus, $\mathbb{E}[\tilde{T}(n)]$ is upper bounded as follows:

$$\mathbb{E}[\tilde{T}(n)] \leq \sum_{t=1}^{n} \mathbb{P}\Big(\frac{\log(t)}{T(t-1)} > \frac{(\tau-\mu)^2}{2\beta}\Big) + \frac{\pi^2}{3}. \quad (12)$$

The first term on the right-hand side of (12) is upper bounded by Claim 4 with $\epsilon = \frac{(\tau-\mu)^2}{2\beta}$. Thus the proof follows.

Proof of Claim 2. Suppose neither holds. Then,

$$\begin{aligned} \bar{\mu}_{T(t-1)} + c_{t,T(t-1)} &\leq \mu + 2 \cdot c_{t,T(t-1)} \\ &\leq \mu + 2 \cdot \frac{\tau-\mu}{2} = \tau. \end{aligned}$$

Thus, $I_a(t) \neq k_a$. □

Proof of Claim 3. The decomposition of $T_0^2(n)$ into diagonal and off-diagonal elements and union bound provide the following upper bound:

$$\begin{aligned} T_0^2(n) &\leq \sum_{t=1}^{n} \mathbb{I}_{\{\bar{\mu}_{T(t-1)}(t) \leq d\}} + 2\sum_{t=1}^{n-1}\sum_{s=t+1}^{n} \mathbb{I}_{\{\bar{\mu}_{T(s-1)}(s) \leq d\}} \\ &\leq \sum_{t=1}^{n}\sum_{r=1}^{K_{max}t} \mathbb{I}_{\{\bar{\mu}_r(t) \leq d\}} + 2\sum_{t=1}^{n-1}\sum_{s=t+1}^{n} \mathbb{I}_{\{\bar{\mu}_r(s) \leq d\}}. \end{aligned}$$

Taking the expectation, and applying Hoeffding-Chernoff Bound,

$$\begin{aligned} \mathbb{E}T_0^2(n) &\leq K_{max}\sum_{t=1}^{n} t^{1-\beta} + 2K_{max}\sum_{t=1}^{n-1}\sum_{s=t+1}^{n} s^{1-\beta} \\ &\leq K_{max}\sum_{t=1}^{n} t^{1-\beta} + 2K_{max} \cdot \sum_{t=1}^{n} t^{2-\beta} \\ &\leq K_{max}(\frac{\pi^2}{6} + \frac{\pi^2}{3}) = K_{max}\frac{\pi^2}{2}. \end{aligned}$$

□

Proof of Claim 4. Fix $\epsilon > 0$. Note that

$$\begin{aligned} T(n) &= n - T_0(n) + T_2(n) \\ &\geq n - T_0(n). \end{aligned}$$

Therefore,

$$\mathbb{P}\Big(\frac{\log(t+1)}{T(t)} > \epsilon\Big) \leq \mathbb{P}\Big(T_0(t) > t - \frac{\log(t+1)}{\epsilon}\Big). \quad (13)$$

If $t - \frac{\log(t+1)}{\epsilon} > 0$, Markov Inequality applied to the RHS of (13) implies the following:

$$\mathbb{P}\Big(T_0(t) > t - \frac{\log(t+1)}{\epsilon}\Big) \leq \frac{\mathbb{E}T_0^2(t)}{(t - \frac{\log(t+1)}{\epsilon})^2}. \quad (14)$$

Let $t_\epsilon = \inf\{t : t - \frac{\log(t+1)}{\epsilon} > 0\}$. Then,

$$\begin{aligned} \sum_{t=0}^{\infty} \mathbb{P}\Big(\frac{\log(t+1)}{T(t)} > \epsilon\Big) &\leq t_\epsilon + K_{max}\frac{\pi^2}{2}\sum_{t=1}^{\infty}\frac{1}{(t - \frac{\log(t+1)}{\epsilon})^2} \\ &= M(\epsilon) \\ &< \infty, \end{aligned}$$

where the first line above follows from Claim 3. Thus the proof follows.

□

REFERENCES

[1] P. Kyasanur, N. H. Vaidya, "Capacity of multi-channel wireless networks: impact of number of channels and interfaces" Proceedings of the 11th Annual International Conference on Mobile Computing and Networking. ACM, 2005.

[2] T. S. Rappaport et al. "Millimeter Wave Wireless Communications", Pearson Education, 2014.

[3] S. Cayci and A. Eryilmaz, "On the Multi-Channel Capacity Gains of Millimeter-Wave Communication," 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, 2016, pp. 1-6.

[4] P. Bahl, A. Adya, J. Padhye, A. Wolman, Reconsidering Wireless Systems with Multiple Radios, ACM Computing Communication Review, July 2004.

[5] D. Tse, P. Viswanath, "Fundamentals of Wireless Communication", Cambridge University Press, 2005.

[6] Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS), "Deliverable D1.4 METIS Channel Models", Document Number: ICT-317669-METIS/D1.4.

[7] M. R. Akdeniz et al., "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation", in IEEE Journal on Selected Areas in Communications, vol. 32, no. 6, pp. 1164-1179, June 2014.

[8] S. Rangan, T. S. Rappaport and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges", in Proceedings of the IEEE, vol. 102, no. 3, pp. 366-385, March 2014.

[9] M. J. Neely, E. Modiano, C. E. Rohrs, "Dynamic power allocation and routing for time varying wireless networks", in Proc. IEEE International Conference on Computer Communications (INFOCOM), San Francisco, CA, April 2003.

[10] M. J. Neely, E. Modiano, C. Li., "Fairness and optimal stochastic control for heterogeneous networks" , in Proc. IEEE International Conference on Computer Communications (INFOCOM), Miami, FL, March 2005.

[11] A. Eryilmaz, R. Srikant, J. R. Perkins, "Stable scheduling policies for fading wireless channels", IEEE/ACM Transactions on Networking, 13:411425, April 2005.

[12] L. Georgiadis, M. J. Neely, L. Tassiulas. "Resource allocation and cross-layer control in wireless networks", Foundations and Trends in Networking, vol. 1, no. 1, pp. 1-149, 2006.

[13] J. W. Lee, R. R. Mazumdar, N. B. Shroff, "Downlink power allocation for multi-class cdma wireless networks", Proc. IEEE INFOCOM, 2002.

[14] H. Gangammanavar and A. Eryilmaz, "Dynamic coding and rate-control for serving deadline-constrained traffic over fading channels", in Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on, pages 1788 1792, June 2010.

[15] R. Aggarwal, M. Assaad, C. E. Koksal, P. Schniter. "OFDMA Downlink Resource Allocation via ARQ Feedback", in 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, pages 14931497, Nov 2009.

[16] R. Aggarwal, P. Schniter, C. E. Koksal, "Rate Adaptation via Link-Layer Feedback for Goodput Maximization over a Time-Varying Channel", IEEE Transactions on Wireless Communications, 8(8):42764285, August 2009.

[17] S. Bubeck, N. Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems", Foundations and Trends in Machine Learning, Vol. 5, No. 1, pp. 1122, 2012.

[18] C. Tekin, M. Liu, "Online Learning Methods for Networking", Foundations and Trends in Networking: Vol. 8: No. 4, pp 281-409, 2015 http://dx.doi.org/10.1561/1300000050

[19] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time analysis of the multiarmed bandit problem", Machine Learning Vol. 47, no. 2-3, pp. 235-256, 2002.

[20] S. Krishnasamy, R. Sen, R. Johari, S. Shakkottai, "Regret of Queueing Bandits", CoRR, abs/1604.06377, 2016.

[21] S. Agrawal, N. Goyal, "Analysis of Thompson Sampling for the Multi-armed Bandit Problem", Proceedings of 25th Annual Conference on Learning Theory, vol. 23, pp. 1-39, 2012.

[22] A. Gopalan, S. Mannor, Y. Mansour, "Thompson sampling for complex bandit problems." arXiv preprint arXiv:1311.0466 (2013).