Learning to Control Renewal Processes with Bandit Feedback

Semih Cayci
ECE, The Ohio State University
Columbus, OH
cayci.1@osu.edu

Atilla Eryilmaz
ECE, The Ohio State University
Columbus, OH
eryilmaz.2@osu.edu

R. Srikant
CSL and ECE, UIUC
rsrikant@uiuc.edu

ABSTRACT

We consider a bandit problem with K task types from which the controller activates one task at a time. Each task takes a random and possibly heavy-tailed completion time, and a reward is obtained only after the task is completed. The task types are independent from each other, and have distinct and unknown distributions for completion time and reward. For a given time horizon τ, the goal of the controller is to schedule tasks adaptively so as to maximize the reward collected until τ expires. In addition, we allow the controller to interrupt a task and initiate a new one. In addition to the traditional exploration-exploitation dilemma, this interrupt mechanism introduces a new one: should the controller complete the task and get the reward, or interrupt the task for a possibly shorter and more rewarding alternative? We show that for all heavy-tailed and some light-tailed completion time distributions, this interruption mechanism improves the reward linearly over time. From a learning perspective, the interrupt mechanism necessitates implicitly learning statistics beyond the mean from truncated observations. For this purpose, we propose a robust learning algorithm named `UCB-BwI` based on the median-of-means estimator for possibly heavy-tailed reward and completion time distributions. We show that, in a K-armed bandit setting with an arbitrary set of L possible interrupt times, `UCB-BwI` achieves $O(K\log(\tau)+KL)$ regret. We also prove that the regret under any admissible policy is $\Omega(K\log(\tau))$, which implies that `UCB-BwI` is order optimal.

CCS CONCEPTS

• **Theory of computation → Online learning theory**.

KEYWORDS

multi-armed bandits, online learning, renewal theory, stochastic knapsack, heavy-tailed distributions, online scheduling

ACM Reference Format:
Semih Cayci, Atilla Eryilmaz, and R. Srikant. 2019. Learning to Control Renewal Processes with Bandit Feedback. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '19 Abstracts), June 24–28, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3309697.3331515

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGMETRICS '19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6678-6/19/06.
https://doi.org/10.1145/3309697.3331515

1 INTRODUCTION

In many real life problems, a server processes tasks with random completion times that are unknown in advance, and the controller schedules these tasks so as to maximize the number of task completions in a given time interval. The service time distribution is often heavy-tailed in many economic, social and technological systems, which implies that the mean residual time to complete a task grows over time [2, 6]. As a consequence, in addition to the conventional exploration-exploitation dilemma considered in [1, 7], the controller faces with a new dilemma: after initiating a task, should it wait until completion and gather the reward, or make a new decision that could possibly serve faster at the expense of rejecting the reward and wasting the time already spent? As we show in this work, this interruption mechanism becomes crucial for optimal performance. In this paper, we model this reward maximization problem as a continuous-time multi-armed bandit (MAB) problem with possibly heavy-tailed reward and completion time distributions, in which the controller has the option to interrupt a task at any time and make a new decision. The applications of this model include task scheduling, adaptive routing and online learning for optimal free trial duration in advertising.

2 PROBLEM FORMULATION

We consider a set of K statistically independent task types (or arms), denoted by $\mathcal{K}=\{1,2,\dots,K\}$. Each arm corresponds to a stochastic process $\{(X_n^{(k)},R_n^{(k)}),\ n\geq 1\}$. If arm k is activated (i.e., a task of type k is initiated) at the time of n-th decision, it takes a random *completion time* $X_n^{(k)}>0$ to obtain the *reward* $R_n^{(k)}\geq 0$ at the end. For a given time horizon $\tau>0$, the sequential decision-making continues until the time expires. Both $X_n^{(k)}$ and $R_n^{(k)}$ are unknown to the controller when the decision is made. The stochastic process $\{(X_n^{(k)},R_n^{(k)})\}$ corresponding to arm k is independent and identically distributed (iid) over n, therefore it is a renewal reward process. We assume that $X_n^{(k)}$ and $R_n^{(k)}$ are independent random variables, and the following moment condition is satisfied by all arms:

$$\max\{\mathbb{E}[(X_1^{(k)})^{1+\gamma}],\mathbb{E}[(R_1^{(k)})^{1+\gamma}]\}<\infty,\ \forall k\in\mathcal{K}, \tag{1}$$

for some $\gamma\in(0,1]$. Therefore, the model includes heavy-tailed reward and completion time distributions.

In this problem, the controller has to make two decisions: the task type and the interrupt time. Let $\mathcal{B}\subset\mathbb{R}_+$ be the set of interrupt times that will be specified later. A policy $\pi=\{\pi_n\}_{n=1}^{\infty}$ consists of two parts: $\pi_n=(I_n,B_n^{(I_n)})\in\mathcal{K}\times\mathcal{B}$. A decision $\pi_n=(k,b)$ implies that a task of type k is activated at the time of n-th decision, and an interrupt time of b time units from the activation time is declared. For a control $\pi_n=(k,b)$, the completion time of a task is $(X_n^{(k)}\wedge b)$, the reward is $R_n^{(k)}\mathbb{I}_{\{X_n^{(k)}\leq b\}}$. Therefore, the problem at

hand is an exploration-exploitation problem in which the learning is conducted via right-censored feedback.

For a given admissible policy π, the counting process $N_\pi(\tau)$ is the total number of completed tasks in $[0,\tau]$. Then, the cumulative reward under π is as follows:

$$Rew_\pi(\tau) = \sum_{n=1}^{N_\pi(\tau)} \sum_{(k,b)\in\mathcal{K}\times\mathcal{B}} \mathbb{I}_{\{\pi_n=(k,b)\}} R_n^{(k)} \mathbb{I}_{\{X_n^{(k)}\le b\}}. \quad (2)$$

The objective in this paper is to design online learning algorithms that maximize the expected cumulative reward, or equivalently minimize regret, which is defined as follows:

$$\overline{Reg}_\pi(\tau) = \max_{\pi'} \mathbb{E}[Rew_{\pi'}(\tau)] - \mathbb{E}[Rew_\pi(\tau)]. \quad (3)$$

where the maximization is over the set of all admissible policies. The regret of a policy π is the loss suffered due to suboptimal decisions in both arm and interrupt time selection.

3 OPTIMAL POLICY WITH KNOWN STATISTICS

Note that this is an extension of the stochastic knapsack problem, and the solution is NP-hard even if all distributions are known [4]. As an approximation, consider the following policy π^*: at each epoch, make the decision (k,b) that maximizes the reward rate $r^{(k)}(b)$, which is defined as follows:

$$r^{(k)}(b) = \frac{\mathbb{E}[R_1^{(k)}\mathbb{I}_{\{X_1^{(k)}\le b\}}]}{\mathbb{E}[X_1^{(k)}\wedge b]}. \quad (4)$$

We have the following optimality gap for the optimal static policy π^*:

$$\max_\pi \mathbb{E}[Rew_\pi(\tau)] - \mathbb{E}[Rew_{\pi^*}(\tau)] \le 2\max_{k\in\mathcal{K}} \mathbb{E}[R_1^{(k)}],\ \forall \tau > 0.$$

Consequently, π^* is asymptotically optimal as $\tau\to\infty$.

The optimal interrupt time for arm k is the following:

$$b_k^* = \sup\{b\in\mathcal{B} : r^{(k)}(b) \ge r^{(k)}(b'), \forall b'\in\mathcal{B}\}. \quad (5)$$

We show that interrupting a task before its completion is optimal, i.e., $b_k^* < \infty$ if and only if $\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] > \mathbb{E}[X_1^{(k)}]$ holds for some $b > 0$. Note that all heavy-tailed and some light-tailed completion time distributions (such as hyperexponential) satisfy this condition, and thus require a finite optimal interrupt time.

In the next section, we propose a UCB-type algorithm for learning the optimal (k, b_k^*) pair.

4 ALGORITHM DESIGN

We consider a finite set of interrupt times $\mathcal{B} = \{b_1, b_2, \dots, b_L\}$ such that $b_1 \le b_2 \le \dots \le b_L = \infty$ without loss of generality. In the absence of the reward rates $r^{(k)}(b)$, we propose an algorithm named UCB-BwI to learn (k, b_l) pair with the highest reward rate.

4.1 Information Structure

The problem has a specific information structure as follows. For an arm k, if an interrupt decision $b_{l'} \ge b_l$ is made, then the feedback gives full information for the decision (k, b_l) as well. If we denote the number of (k, b_l) decisions at the end of n-th epoch by $T_l^{(k)}(n)$, the effective sample size is $\overline{T}_l^{(k)}(n) = \sum_{j\ge l} T_j^{(k)}(n)$, which might be significantly larger than $T_l^{(k)}(n)$.

4.2 UCB-BwI Algorithm and Regret Bound

Since $X_n^{(k)}$ and $R_n^{(k)}$ can be potentially heavy-tailed, we construct a UCB-type policy based on the median-of-means estimator [3]. At epoch $(n+1)$, given that there are $s_{k,l} = \overline{T}_l^{(k)}(n)$ observations for (k, b_l), we compute median-of-means estimators $\overline{U}_{n,s_{k,l}}^{(k)}(b_l)$ and $\overline{V}_{n,s_{k,l}}^{(k)}(b_l)$ for $(X_i^{(k)}\wedge b_l)$ and $(R_i^{(k)}\mathbb{I}_{\{X_i^{(k)}\le b_l\}})$, respectively. Then, the UCB-BwI Algorithm makes a decision as follows:

$$(I_{n+1}, B_{n+1}^{(I_{n+1})}) \in \arg\max_{(k\times b_l)\in\mathcal{K}\times\mathcal{B}} \left\{ \frac{\overline{V}_{n,s_{k,l}}^{(k)}(b_l)}{\overline{U}_{n,s_{k,l}}^{(k)}(b_l)} + \frac{(1+r)\epsilon_{n,s_{k,l}}}{\mu + \epsilon_{n,s_{k,l}}} \right\},$$

where $r \ge r^{(k)}(b)$, $\mu \le \mathbb{E}[X_1^{(k)}\wedge b]$ for all (k,b), and

$$\epsilon_{n,s} = \beta\left[\frac{\log\left(2e^{\frac{1}{8}}(n+1)^4\right)}{s}\right]^{\frac{\gamma}{1+\gamma}},$$

for some $\beta > 0$.

In the following, we present a distribution-dependent regret upper bound for UCB-BwI.

Theorem 4.1 (Regret Upper Bound for UCB-BwI). *The regret under* UCB-BwI *satisfies the following for all* $\tau > 0$*:*

$$\overline{Reg}_{\pi^{\text{BwI}}}(\tau) \le \sum_{k:d^{(k)}>0} \left[C^{(k)}\log\left(\frac{\tau}{\mu}\right) + O\left(\frac{L}{\left(d_{min}^{(k)}\right)^{\frac{1}{\gamma}}}\right)\right] + O(KL),$$

where

$$C^{(k)} = C_1\left(\frac{1+r}{\mu}\right)^{\frac{\gamma+1}{\gamma}}\left[\left(\frac{1}{d_{min}^{(k)}}\right)^{\frac{1}{\gamma}} + \left(\frac{1}{d^{(k)}}\right)^{\frac{1}{\gamma}}\right],$$

for some constant $C_1 > 0$, $d_{min}^{(k)} = \min_{l:b_l\ne b_k^*}\left\{r^{(k)}(b_k^*) - r^{(k)}(b_l)\right\}$ *and* $d^{(k)} = \max_{(k',b)} r^{(k')}(b) - r^{(k)}(b_k^*)$.

According to Theorem 4.1, the regret grows at a rate $O(K\log(\tau) + KL)$. As a result of the specific information structure, the coefficient of the time-dependent term in the regret, $C^{(k)}$, is independent of $L = |\mathcal{B}|$.

By using a similar approach to [5], we also show that the regret has a lower bound $\Omega(K\log(\tau))$. Together with Theorem 4.1, this implies that UCB-BwI is order optimal in K, L and τ.

REFERENCES

[1] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. 2013. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 207–216.

[2] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207.

[3] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. 2013. Bandits with heavy tail. *IEEE Transactions on Information Theory* 59, 11 (2013), 7711–7717.

[4] Brian C Dean, Michel X Goemans, and Jan Vondrdk. 2004. Approximating the stochastic knapsack problem: The benefit of adaptivity. In *45th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, 208–217.

[5] Tze Leung Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6, 1 (1985), 4–22.

[6] Jayakrishnan Nair, Adam Wierman, and Bert Zwart. 2013. The fundamentals of heavy-tails: properties, emergence, and identification. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 41. ACM, 387–388.

[7] Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2015. Thompson Sampling for Budgeted Multi-Armed Bandits.. In *IJCAI*. 3960–3966.