Learning to Control Renewal Processes with Bandit Feedback

SEMIH CAYCI, ECE, The Ohio State University
ATILLA ERYILMAZ, ECE, The Ohio State University
R. SRIKANT, CSL and ECE, University of Illinois at Urbana-Champaign

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. Applications of this model include server scheduling, optimal free sampling strategies in advertising and adaptive content selection. From a learning perspective, the interrupt mechanism necessitates learning the whole arm distribution from truncated observations. For this purpose, we propose a robust learning algorithm named `UCB-BwI` based on 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; stochastic scheduling

ACM Reference Format:
Semih Cayci, Atilla Eryilmaz, and R. Srikant. 2019. Learning to Control Renewal Processes with Bandit Feedback. *Proc. ACM Meas. Anal. Comput. Syst.* 3, 2, Article 43 (June 2019), 32 pages. https://doi.org/10.1145/3326158

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 cumulative reward, e.g., the number of task completions, in a given time interval. In many social, economic and technological systems, the service time distribution is often heavy-tailed, which implies that the mean residual time to complete a task grows over time [4, 30]. As a consequence, in addition to the conventional exploration-exploitation dilemma, 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

Authors' addresses: Semih Cayci, ECE, The Ohio State University, Columbus, OH, 43210, cayci.1@osu.edu; Atilla Eryilmaz, ECE, The Ohio State University, Columbus, OH, 43210, eryilmaz.2@osu.edu; R. Srikant, CSL and ECE, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, rsrikant@uiuc.edu.

2476-1249/2019/6-ART43 $15.00
https://doi.org/10.1145/3326158

serve faster at the expense of rejecting the reward and wasting the time already spent? As we will show in this work, this interruption mechanism becomes crucial in maximizing the cumulative reward in a given time interval. We model this problem as a continuous-time multi-armed bandit (MAB) problem in which the controller has the option to interrupt a task any time and make a new decision. Applications of this framework include task scheduling in a multi-server system (such as a cloud computing system or a call center), adaptive routing in networks and optimal free-trial strategy in marketing digital information goods (see Section 3).

In this paper, we consider a continuous-time bandit setting where the controller can make a decision at any time. Each arm (or task type) is modeled as a renewal reward process, where each task corresponds to a renewal epoch from the selected task type. Each task takes a random completion time, which is unknown to the controller in advance, and a reward is obtained only after the task is completed. The objective of the controller is to maximize the cumulative reward until it runs out of time. Unlike the existing MAB models where the controller makes a decision only after a task is completed [3, 34], we allow the controller to interrupt an ongoing task at any time, and initiate a new task of possibly different type at the expense of rejecting the reward of the last task. As we show in this paper, this interrupt mechanism becomes essential when the completion time distribution is heavy-tailed.

From a learning perspective, implications of an interrupt mechanism are two-fold:

(1) The controller has to learn the *whole* distribution unlike the existing MAB models in which learning only the first-order statistics is sufficient.
(2) This is an exploration-exploitation problem with possibly right-censored observations. In order to see this, consider an interrupted task. Since the task is not finalized, the reward and completion time realizations are not received, and the controller only receives the feedback that the completion time exceeded a threshold. Also, the observation for a specific interrupt time provides information about different interrupt times, therefore there is a structure in this problem. As we will see, exploiting this structure in algorithm design becomes crucial in optimizing the performance.

A good algorithm should address these considerations, and it must be statistically robust: in most applications, the arms have diverse and potentially heavy-tailed completion time and reward distributions, and an algorithm must be able to perform well under this heterogeneity. The objective in this paper is to propose provably efficient and robust learning solutions to this problem.

1.1 Related Work

Multi-armed bandits have been the primary model for sequential decision problems that involve exploration-exploitation trade-off in learning, which is said to be "a conflict evident in all human action" by Whittle in [33]. As a result of this universality, there is a broad list of applications of MAB models ranging from routing in networks to dynamic pricing. For excellent surveys in stochastic bandit theory, we refer to [6] and [7].

Continuous-time bandits have been considered in different contexts. In [23] and [27], the problem is explored from a classical bandit perspective, and Gittins index policies are proposed to maximize cumulative reward. In [3, 19, 34, 35], continuous-time bandits are investigated from a frequentist perspective. In all of these works, the completion times are assumed to be $[0, 1]$-valued random variables. Also, the controller makes a decision only when a task is completed. However, in many applications, especially in the ones that involve "human action", the completion times and rewards naturally have heavy-tailed distributions [4], and thus an interrupt mechanism, i.e., a *true* continuous-time bandit setting is required for optimal performance. Thus, the existing approaches fall short to solve our problem, and the proposed algorithms cannot achieve sublinear regret.

Bandits with heavy-tailed distributions were considered in [8, 26]. In both of these works, the authors extend the existing results for sub-Gaussian distributions to the case where the *reward* distributions are heavy-tailed. The setting in these works is still the classical discrete-time MAB setting, and the challenge is to find robust estimators for heavy-tailed reward distributions. In our setting, the duration of a task, which is subject to a constraint, is heavy-tailed, and thus new dynamics are introduced to the problem.

1.2 Contributions

Our main contributions in this work are the following:

- We introduce the bandits with interrupts (BwI) framework. We identify the conditions on the arm distributions under which interrupting an ongoing task is optimal. By using tools from renewal theory, we determine an asymptotically optimal and tractable policy given arm statistics (see Sections 2 and 4).
- We present concentration inequalities for renewal reward processes, which are fundamental in algorithm design and analysis in our problem among many other applications (see Section 5.1).
- We propose a UCB-type learning algorithm for the BwI problem, called the UCB-BwI Algorithm, which is non-parametric in the sense that it requires no other assumptions on arm statistics than the usual moment assumptions (see Section 5.3). Then, in Section 6, we prove that UCB-BwI achieves $O(K\log(\tau) + KL)$ over a time interval τ with L being the cardinality of the set of possible interrupt times. Moreover, we show that UCB-BwI is order-optimal in K, L and τ by showing that the regret under any admissible policy is $\Omega(K\log\tau)$ as $\tau \to \infty$.

1.3 Notation

Throughout this paper, we denote the minimum of two numbers as $\min\{a, b\} = a \wedge b$, and maximum of two numbers as $\max\{a, b\} = a \vee b$ interchangeably for any $a, b \in \mathbb{R}$. The cardinality of a set A is denoted as $|A|$, and the complement of a set A is denoted as A^c. For any event E, $\mathbb{I}_E$ denotes the indicator function.

2 PROBLEM FORMULATION

We consider a set of K statistically independent task types (or arms), denoted by $\mathcal{K} = \{1, 2, \ldots, K\}$. To follow the bandit terminology, we will use "arms" synonymously with "task types". 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* $X_n^{(k)}$ to obtain the *reward* $R_n^{(k)}$ at the end. For a given time horizon $\tau > 0$, the sequential decision-making continues until the time horizon 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 $X_n^{(k)} > 0$ and $R_n^{(k)} \geq 0$ 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, this model includes heavy-tailed reward and completion time distributions.

In the BwI problem, the goal is to maximize the cumulative reward collected within a given time interval $[0, \tau]$. Consequently, the completion time of a task is as important as the reward it yields. As a distinctive feature of our model, we give the controller the option to interrupt an ongoing task if it takes too much time. If a task is interrupted, the reward of that task is rejected, and a new task is initiated immediately for a possibly more rewarding alternative. This introduces a new control dimension to our model, which is not present in the existing bandit models. As it will be seen in the next section, this control dimension is vital for optimal performance in a broad class of arm distributions.

Formally, in the BwI 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 $R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \leq b\}}$, and therefore a stochastic feedback Y_n is obtained as follows:

$$Y_n(k, b) = \Big(\mathbb{I}_{\{X_n^{(k)} \leq b\}},\ X_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \leq b\}},\ R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \leq b\}}\Big). \tag{2}$$

In words, the knowledge about $(X_n^{(k)}, R_n^{(k)})$ is obtained only if the task is completed before b, and only $\mathbb{I}_{\{X_n^{(k)} \leq b\}}$ is obtained otherwise. Therefore, the problem at hand is an exploration-exploitation problem in which the learning is conducted via right-censored feedback. Now we formally define an admissible learning policy for this setting.

Definition 2.1 (Admissible Policy). Let $\mathcal{F}_n = \sigma(Y_1, Y_2, \ldots, Y_{n-1})$ be the history of the received feedback up to n-th decision, where $\sigma(Y)$ denotes the sigma-field of a random variable Y. We call a policy $\pi = \{\pi_n\}_{n=1}^{\infty}$ admissible if $\pi_n \in \mathcal{F}_n$ for all n.

We call the period between n-th and $(n+1)$-th decisions n-th epoch. In the traditional MAB models, the number of epochs is a given deterministic quantity, which is equal to the time horizon. However, in the BwI problem, the number of epochs until the time expires is random. In the following, we define the number of epochs as a counting process.

Definition 2.2 (Counting Process). For a given admissible policy π, let

$$S_n^{\pi} = \sum_{s=1}^{n} \sum_{(k,b) \in \mathcal{K} \times \mathcal{B}} \mathbb{I}_{\{\pi_s = (k,b)\}} \big(X_s^{(k)} \wedge b\big), \tag{3}$$

be the total time spent at the end of the n-th epoch. Given a time horizon τ, the counting process $N_\pi(\tau)$, which denotes the total number of completed tasks in $[0, \tau]$ is defined as follows:

$$N_\pi(\tau) = \sup\{n : S_n^{\pi} \leq \tau\}. \tag{4}$$

Note that $\nu_\pi(\tau) = N_\pi(\tau) + 1$ is known as the first passage time, and is a stopping time under any admissible policy π [16, 18].

Definition 2.3 (Cumulative Reward under a Policy). Let π be an admissible policy. Then, the cumulative reward under π is as follows:

$$Rew_\pi(\tau) = \sum_{n=1}^{N_\pi(\tau)} \sum_{(k,b) \in \mathcal{K} \times \mathcal{B}} \mathbb{I}_{\{\pi_n = (k,b)\}} R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \leq b\}}, \tag{5}$$

where $N_\pi(\tau)$ is the counting process in (4).

In Figure 1, we illustrate a sample path from the BwI process.

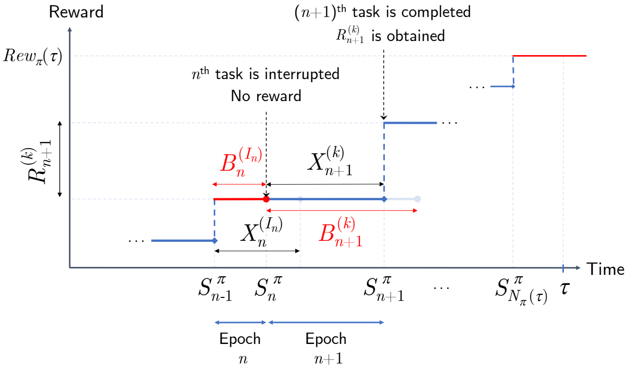

Fig. 1. Illustration of a sample path for the BwI process. The completion time $X_n^{(I_n)}$ of the n-th task exceeds the interrupt time $B_n^{(I_n)}$, so the task is interrupted. In the $(n+1)$-th epoch, arm k is chosen, and the task is completed before the interrupt time, yielding a reward of $R_{n+1}^{(k)}$.

Definition 2.4 (Regret). For a given time horizon $\tau > 0$, the optimal policy maximizes the cumulative reward within $[0, \tau]$:

$$\pi^{\text{opt}} = \arg\max_{\pi} \mathbb{E}\Big[Rew_\pi(\tau)\Big], \tag{6}$$

where the maximization is over the set of all admissible policies. The objective in this paper is to design online learning algorithms that have a good competitive performance with respect to π^{opt}. The performance metric for this objective is the regret, which is defined as follows:

$$\overline{Reg}_\pi(\tau) = \mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)] - \mathbb{E}[Rew_\pi(\tau)]. \tag{7}$$

The regret of a policy π is the loss suffered due to suboptimal decisions in both arm and interrupt time selection.

In the next section, we formulate some problems that can be solved within the BwI framework.

3 NOTABLE EXAMPLES

In many technological, economic and social systems, resource allocation is performed over alternative tasks with randomly varying resource consumption. As such, the controller has to track and possibly interrupt the resource consumption of each activated task for efficient utilization of the budget. The BwI framework that we introduce and investigate in the paper is aimed at forming the foundation for developing effective learning solutions for optimal resource allocation in such problems. In the following, we note some notable examples, for which our proposed BwI framework and design can be utilized.

(1) Adaptive Routing in Communications: In the first example, we consider an adaptive routing problem with the objective of throughput maximization. In a broad class of communication systems, the transmission times are unknown to the controller at the time of scheduling, and learned via ACK/NACK type feedback [29]. For these systems equipped with ARQ control, the packet transmission times might possibly follow a heavy-tailed distribution depending on the channel statistics [21, 32]. In the absence of any statistical knowledge about the channels, the goal of the controller is to learn the best channel based on the feedback so as to maximize the throughput.

As an instance of the systems described above, consider a simplistic point-to-point communication scenario with K parallel and independent channels with ARQ control to transmit packets successfully. If channel k is scheduled to transmit n-th packet, it takes $X_n^{(k)}$ units of time to complete the transmission. The statistical properties of the channels are unknown, and learned by ACK/NACK type feedback. For a given time horizon τ, the goal of the controller is to maximize throughput, i.e., transmit as many packets as possible in the limited time interval $[0, \tau]$. Since the time is limited, at each transmission, the controller determines an interrupt time of $B_n^{(k)}$ time units to avoid extremely long transmission times. If the ACK signal does not arrive within $B_n^{(k)}$ time units after the transmission, then the session is interrupted and a new session is initiated with a possibly different channel. This adaptive scheduling procedure can be posed as an optimization problem as follows:

$$\max_{\pi} \mathbb{E}\Big[\sum_{n=1}^{N_\pi(\tau)} \mathbb{I}_{\{X_n^{(k)} \leq B_n^{(k)}\}}\Big].$$

It is shown in [21, 32] that the retransmission protocol might lead to heavy-tailed completion times $X_n^{(k)}$ even if all the system components are light-tailed. As we will show in the next section, there is a finite optimal interrupt time if the transmission time has a heavy-tailed distribution. In the absence of channel statistics, the controller must learn the whole distribution of $X_n^{(k)}$ via ACK/NACK type feedback to choose the optimal channel and interrupt time. Thus, this adaptive routing problem falls into the BwI framework.

Akin to the classical multi-armed bandit framework guiding the solution of many learning-optimization problems, our BwI framework can form the foundation for the solution of a large class of stochastic scheduling problems by extending the model to specific bandit scenarios. For example, in a realistic communication system, if the transmission time is too long for a channel, then it will possibly yield a long transmission time if it is scheduled immediately for retransmission. In order to model such scenarios, the BwI model can be extended to a sleeping [24] or Markovian [1] bandit setting. On the other hand, the controller might transmit a new packet without waiting for the ACK signal of the previous transmission, which can be modeled by a delayed feedback extension of the BwI framework [22].

(2) Task Scheduling in Data Centers: Consider a computing system with a single processor and K different user types with distinct usage characteristics. Let the task length of a user of type k be denoted as $X_n^{(k)}$, and a payoff $R_n^{(k)}$ is received at the end of the service. For example, if the goal is to maximize throughput, there is a unit reward, i.e., $R_n^{(k)} = 1$, after each service completion. For a given time horizon τ, which is the duration of the busy period of the processor, the goal is to maximize the total reward in $[0, \tau]$ by scheduling tasks. Empirical studies in a variety of private enterprise, campus and cloud data centers indicate that the task sizes and arrivals exhibit highly heavy-tailed characteristics in such systems [5, 20], therefore the sequential scheduling problem subject to time constraints can be solved within the BwI framework.

(3) Optimal Free Trial Strategy: Advertising through free samples is ubiquitous in markets for a big variety of goods such as games, software-as-a-service (SaaS) applications or material goods. In such scenarios, there are multiple types of target populations with distinct utilization preferences. In order to make use of the budget efficiently while maximizing the revenue, user characterization and resource allocation must be performed in an optimal and data-driven manner. In a simplistic market model, for a type-k user, the free-to-paid conversion takes $X_n^{(k)}$ time (or resource), and a reward of $R_n^{(k)}$ is obtained if the user buys the product. Depending on the users' behavioral preferences such as free riding, $X_n^{(k)}$ might follow a heavy-tailed distribution. For empirical studies, we refer to [15, 31]. This type of heavy-tailed behavior necessitates limiting the amount of free sampling for some user types. In order to maximize the number of free-to-paid conversions within an allowed campaign resource budget τ, the controller has to find the optimal target population along with

an optimal length of free-trials. Thus, this problem can be formulated as an instance of the BwI setting. We note that it is possible to address multiple customer types simultaneously by using a combinatorial bandit [11] extension of the BwI framework, which is a direction of future research.

4 OPTIMAL POLICY WITH KNOWN STATISTICS

In this section, we investigate the characteristics of the optimal policy for the problem (6) when all arm statistics are given as oracles. Note that (6) is a stochastic knapsack problem, and the solution is NP-hard even if all distributions are known [14], which makes the learning and competitive analysis intractable. In order to solve this problem, we focus on approximation algorithms for the BwI problem. In the following, we propose a simple static policy, and show that the optimality gap it is $O(1)$ as $\tau \to \infty$, which implies an effective finite time performance and asymptotic optimality.

We begin by formally defining the concept of static policies.

Definition 4.1 (Static Policy). Let $k \in \mathcal{K}$ and $b \in \mathcal{B}$. For any $\tau > 0$, the static policy $\pi^{(k)}(b)$ pulls the arm k with a fixed interrupt time $b > 0$ consistently until the time expires, i.e., $(I_n, B_n^{(I_n)}) = (k, b)$ for all n under $\pi^{(k)}(b)$.

Note that for any $(k,b) \in \mathcal{K} \times \mathcal{B}$, the observed stochastic process under the static policy $\pi^{(k)}(b)$ is iid over n, thus it is a renewal reward process. As a fundamental result of the renewal theory, the time average reward per unit time under $\pi^{(k)}(b)$ converges to a positive constant as $\tau \to \infty$:

$$\lim_{\tau \to \infty} \frac{\mathbb{E}[Rew_{\pi^{(k)}(b)}(\tau)]}{\tau} = \frac{\mathbb{E}[R_1^{(k)} \mathbb{I}_{\{X_1^{(k)} \leq b\}}]}{\mathbb{E}[X_1^{(k)} \wedge b]}. \tag{8}$$

Hence, this constant, which is called the reward rate, is the growth rate of the reward over time, and it will be the main quantity of interest throughout the paper.

Definition 4.2 (Renewal Reward Rate). For any $(k,b) \in \mathcal{K} \times \mathcal{B}$, the (renewal) reward rate under the static policy $\pi^{(k)}(b)$ is defined as follows:

$$r^{(k)}(b) = \frac{\mathbb{E}[R_1^{(k)} \mathbb{I}_{\{X_1^{(k)} \leq b\}}]}{\mathbb{E}[X_1^{(k)} \wedge b]}. \tag{9}$$

$r^{(k)}(b)$ is the ensemble average reward per unit time.

Intuitively, if the arm k is chosen, and each task is interrupted at a fixed time $b > 0$ consistently, then the total reward obtained between $[0, \tau]$ is $O\big(\tau \cdot r^{(k)}(b)\big)$.

Let the optimal interrupt time be defined as follows:

$$b_k^* = \sup\{b \in \mathcal{B} : r^{(k)}(b) \geq r^{(k)}(b'), \forall b' \in \mathcal{B}\}. \tag{10}$$

In the following, we investigate the nature of b_k^* depending on the joint distribution of $(X_n^{(k)}, R_n^{(k)})$ in the most general case of $\mathcal{B} = \mathbb{R}_+ \cup \{\infty\}$.

Proposition 4.3 (Optimal Interrupt Time). *Interrupting a task before its completion is optimal, i.e., $b_k^* < \infty$ if and only if the following holds:*

$$\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] > \mathbb{E}[X_1^{(k)}]. \tag{11}$$

for some $b > 0$.

Proof. Take $(k, b) \in \mathcal{K} \times \mathcal{B}$ and let $q = \mathbb{P}(X_1 \leq b)$ and $\mathbf{1}_b = \mathbb{I}_{\{X_1^{(k)} \leq b\}}$. We can write $\mathbb{E}[X_1^{(k)} \wedge b] = \mathbb{E}[X_1^{(k)}\mathbf{1}_b] + b(1-q)$. Then,

$$r^{(k)}(b) - \frac{\mathbb{E}[R_1^{(k)}]}{\mathbb{E}[X_1^{(k)}]} = \mathbb{E}[R_1^{(k)}]\Big(\frac{q}{\mathbb{E}[X_1^{(k)}\mathbf{1}_b] + b(1-q)} - \frac{1}{\mathbb{E}[X_1^{(k)}]}\Big),$$

$$= \frac{\mathbb{E}[X_1^{(k)}(1-\mathbf{1}_b)] - b(1-q) - \mathbb{E}[X_1^{(k)}]}{\mathbb{E}[X_1^{(k)} \wedge b]\mathbb{E}[X_1^{(k)}]/\mathbb{E}[R_1^{(k)}]}.$$

By using the identity

$$\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] = \mathbb{E}[X_1^{(k)}(1-\mathbf{1}_b)] - b(1-q),$$

we can deduce that:

$$r^{(k)}(b) - \frac{\mathbb{E}[R_1^{(k)}]}{\mathbb{E}[X_1^{(k)}]} = \frac{\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] - \mathbb{E}[X_1^{(k)}]}{\mathbb{E}[X_1^{(k)} \wedge b]\mathbb{E}[X_1^{(k)}]/\mathbb{E}[R_1^{(k)}]}.$$

Therefore, $b_k^* < \infty$ if and only if $\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] > \mathbb{E}[X_1^{(k)}]$ for some $b \in (0, \infty)$. □

The quantity $\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b]$ that arises in (11) is called mean residual life, and it quantifies the mean waiting time to get the reward given that the controller has already waited for $b > 0$ time units.

The intuition behind Proposition 4.3 is as follows. At each time instance after a decision, the controller faces a dilemma: continue the ongoing task to get the immediate reward, or interrupt and re-initiate a new task. If there exists $b > 0$ that satisfies (11), then the controller has to wait longer than the average completion time to get the reward, thus interrupting is optimal.

Interruption improves the reward rate for a large class of completion time distributions, including all heavy-tailed and some light-tailed distributions. In the following, we consider some specific classes of distributions, and investigate the behavior of $r^{(k)}(b)$ with respect to the interrupt time b.

Corollary 4.4 (Interrupts for Some Specific Distributions). *Consider the case where $R_n^{(k)}$ is independent of $X_n^{(k)}$.*

1. *If $X_n^{(k)}$ has a heavy-tailed distribution, then interrupting is optimal, i.e., $b_k^* < \infty$.*
2. *If $\mathbb{E}[X_n^{(k)} - b | X_n^{(k)} > b]$ is an increasing function of $b > 0$, then $b_k^* < \infty$.*
3. *If $X_n^{(k)} \sim Exp(\lambda)$ for some $\lambda > 0$, then $r^{(k)}(b) = r^{(k)}(b')$ for all $b, b' > 0$. Exponential distribution is the only completion time distribution with this property.*
4. *If $\mathbb{E}[X_n^{(k)} - b | X_n^{(k)} > b]$ is a monotonically decreasing function of $b > 0$, then $b_k^* = \infty$.*

Remark 1. We make the following observations from Corollary 4.4:

- Tails are important for optimal performance: for two distributions with the same first-order statistics ($\mathbb{E}[X_n^{(k)}], \mathbb{E}[R_n^{(k)}]$), tail statistics might yield very different reward rates $r^{(k)}(b)$ when the interrupt mechanism is employed. If interruption is not employed and $b_k^* < \infty$, then there is a loss that grows linearly in τ.
- Most light-tailed distributions have decreasing mean residual life functions, including Gaussian, uniform, logistic, Laplace and gamma distributions [30]. Part (4) of Corollary 4.4 implies it is optimal to wait until a task is completed for such distributions.
- Exponential distribution serves as a barrier case: interruption does not make a difference if the completion time is exponentially distributed as a consequence of the memoryless property.

- Although the optimal interrupt time is finite for all heavy-tailed, and infinite for most light-tailed completion time distributions, there exist light-tailed distributions for which interruption at a finite time is optimal. Part (2) suggests an example class of completion time distributions that includes hyperexponential distribution.

Proof. (1) is a direct consequence of Property 2.2 in [9]. For (2) and (4), we have $\mathbb{E}[X_1^{(k)}|X_1^{(k)} > 0] = \mathbb{E}[X_1^{(k)}]$ since $X_n^{(k)} > 0$, and therefore monotonicity of the mean residual life implies the corresponding results. For exponential random variables, it holds that $\mathbb{E}[X_1^{(k)} - b | X_1^{(k)} > b] = \mathbb{E}[X_1^{(k)}]$ for all b, which is an alternative form of the memoryless property. □

We illustrate the above results for $X_n^{(k)} \sim Pareto(1, 1.2)$ in Fig. 2 (a) and $X_n^{(k)} \sim Lognormal(1, 2.75)$ in Fig. 2 (b) with $R_n = 1$ for all n.

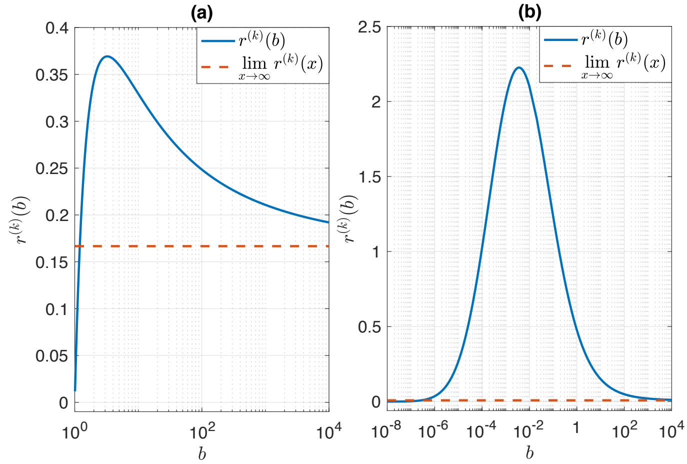

Fig. 2. The renewal reward rate $r^{(k)}(b)$ with respect to b for (a) $X_n^{(k)} \sim Pareto(1, 1.2)$ and (b) $X_n^{(k)} \sim Lognormal(1, 2.75)$. Interruption yields significant gains in the average reward per unit time for heavy-tailed completion time distributions.

Note that interrupting a task at the optimal interrupt time yields significant gains for Pareto distribution, and the impact of interruption becomes drastic in the case of a log-normal distribution with high variability.

From (8), it is seen that a natural approximate algorithm to the optimal policy π^{opt} is the static policy which maximizes $r^{(k)}(b)$ over all (k, b) pairs, which we formally define in the following.

Definition 4.5 (Optimal Static Policy). The optimal static policy, denoted by π^*, makes the following choice at n-th epoch:

$$\pi_n^* = \arg\max_{(k,b)\in\mathcal{K}\times\mathcal{B}} r^{(k)}(b),$$

for all n until the time expires.

In the following, we analyze the performance of this static policy by considering the optimality gap with π^{opt}, and conclude that it achieves an almost-optimal finite-time performance, as well as asymptotic optimality as $\tau \to \infty$.

Proposition 4.6 (Performance of the Optimal Static Policy). *The optimality gap for the optimal static policy π^* bounded for all $\tau > 0$:*

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

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

Proof. Proof of Proposition 4.6 is based on the concept of stopped random walks, and given in Appendix A for the interested reader. □

Remark 2. Proposition 4.6 has a very important consequence: the expected reward under the simple static policy π^* has only a *bounded* gap with the best possible expected reward, which can only be achieved by an NP-hard algorithm for all $\tau > 0$.

In the next section, we propose a UCB-type learning algorithm for the BwI problem.

5 ALGORITHM DESIGN

In this section, we propose a learning algorithm for the BwI problem that converges to the optimal static policy π^* fast enough to yield minimal regret. The learning algorithm has a two-fold objective: (1) learning the optimal interrupt times for all arms, (2) choosing the arm with maximum reward rate. As it was pointed out in Remark 1, interrupt times heavily depend on the tails. Therefore, unlike the traditional bandit models where learning only the first moment suffices, the learning algorithm in this problem must learn the *whole* distribution.

In most real-life problems, the controller has to decide the interrupt time within a discrete set. For example, in the optimal free trial strategy example of Section 3, the duration of free trials is usually measured in terms of days. In digital systems, the processors make decisions discrete in time. Therefore, in this paper, we consider a given finite but arbitrary $\mathcal{B}$ that includes infinite interrupt time (i.e., no interrupt) as an element.

Assumption 1 (Finitely Many Interrupt Times)**.** For $L > 1$, let the set of interrupt times be $\mathcal{B} = \{b_1, b_2, \ldots, b_L\}$ for any user-determined $\{b_i, \ i = 1, 2, \ldots, L\}$. Let

$$0 < b_1 < b_2 < \ldots < b_L = \infty,$$

without loss of generality.

The design strategy will be as follows: each arm-interrupt time pair (k, b) will be a distinct decision, and the objective will be to learn the (k, b) pair with maximum reward rate. In the absence of the arm statistics, an upper confidence bound for the reward rate $r^{(k)}(b)$ will be used as a surrogate. This will be accomplished in three steps:

(1) In Section 5.1, we propose an estimator for the reward rate $r^{(k)}(b)$ based on median-of-means technique, and show that it has an exponential convergence rate by novel concentration inequalities.
(2) In Section 5.2, we examine the specific information structure of the problem to boost the learning rate.
(3) Finally, in Section 5.3, we develop a UCB-type algorithm based on the concentration inequalities we propose, which exploits the information structure of the problem to achieve low regret.

5.1 Concentration Inequalities for Renewal Reward Processes

Recall that for each $(k, b) \in \mathcal{K} \times \mathcal{B}$, the observed stochastic process:

$$\{(X_n^{(k)} \wedge b, R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \leq b\}}), n \geq 1\} \tag{12}$$

is a renewal reward process, and the aim is to find concentration inequalities for $r^{(k)}(b)$ from the first s observations. In the next lemma, we present a concentration result for estimating the ratio of two expected values, e.g., reward rate, by a generic estimator.

Lemma 5.1 (Concentration of Rates). *Let $\{(U_n, V_n),\ n \geq 1\}$ be a sequence of iid vectors with finite mean, and $\bar{\mu}_s = (\bar{U}_s, \bar{V}_s)$ be a generic mean estimator for $(\mathbb{E}[U_1], \mathbb{E}[V_1])$. Let*

$$\Delta_0(\epsilon) = \left(1 + \frac{\mathbb{E}[V_1]}{\mathbb{E}[U_1]}\right)\frac{\epsilon}{\epsilon + \mathbb{E}[U_1]},\ \forall \epsilon > 0. \tag{13}$$

Then, for any p-norm $\|.\|_p, p \geq 1$ defined over $\mathbb{R}^2$, the following inequality is satisfied:

$$\mathbb{P}\left(\frac{\bar{V}_s}{\bar{U}_s} < \frac{\mathbb{E}[V_1]}{\mathbb{E}[U_1]} - \Delta_0(\epsilon)\right) \leq \mathbb{P}(\|\bar{\mu}_s - \mu\|_p > \epsilon), \tag{14}$$

for all $s \geq 1$.

Note that the above concentration inequality yields an upper confidence bound for the reward rate by considering the deviations of U_n and V_n from their means.

Proof. For any $\epsilon > 0$, let the high-probability set $A_s(\epsilon)$ be defined as follows:

$$A_s(\epsilon) = \{\|\bar{\mu}_s - \mu\|_p \leq \epsilon\}.$$

For any outcome $\omega \in A_s(\epsilon)$, it is easy to verify the following:

$$\frac{\bar{V}_s(\omega)}{\bar{U}_s(\omega)} > r - \Delta_0(\epsilon),$$

since the minimum value possible of $\bar{V}_s$ and the maximum value of $\bar{U}_s$ in the set $A_s(\epsilon)$ are $\mathbb{E}[V_1] - \epsilon$ and $\mathbb{E}[U_1] + \epsilon$, respectively. Therefore, the following set inclusion holds:

$$\left\{\frac{\bar{V}_s}{\bar{U}_s} < \frac{\mathbb{E}[V_1]}{\mathbb{E}[U_1]} - \Delta_0(\epsilon)\right\} \subset \left\{\|\bar{\mu}_s - \mu\|_p > \epsilon\right\},$$

which directly implies the result. □

For the observed process (12), the empirical reward rate with s samples is defined as follows:

$$\hat{r}_s^{(k)}(b) = \frac{\sum\limits_{i=1}^{s} R_i^{(k)} \mathbb{I}_{\{X_i^{(k)} \leq b\}}}{\sum\limits_{i=1}^{s} (X_i^{(k)} \wedge b)}. \tag{15}$$

By the fundamental renewal theorem, it is well-known that the empirical reward rate converges to the reward rate almost surely: $\hat{r}_s^{(k)}(b) \longrightarrow r^{(k)}(b)$ almost surely as $s \longrightarrow \infty$ [2, 16]. Therefore, the empirical estimator $\hat{r}_s^{(k)}(b)$ shows up as a natural candidate for estimating $r^{(k)}(b)$. However, for heavy-tailed distributions that satisfy the moment condition (1) for some $\gamma \in (0, 1]$, it can be shown by using Lemma 5.1 and Chebyshev's inequality that the following holds:

$$\mathbb{P}\big(\hat{r}_s^{(k)}(b_L) \leq r^{(k)}(b_L) - \Delta_0(\epsilon)\big) = O\left(\frac{1}{s^\gamma \epsilon^{1+\gamma}}\right),$$

i.e., convergence rate of $\hat{r}_s(b_L)$ is polynomial rather than exponential. Moreover, as it is shown with a lower bound for the convergence rate in [8, 10], this bound is tight. This immediately implies that the empirical reward rate is weak for heavy-tailed distributions, and thus falls short for our application.

For estimating the reward rate, in the following, we present a robust estimator called median-of-means estimator based on [8, 28], which provides exponential convergence rate.

Definition 5.2 (Median-of-Means Estimator). Let $\{(U_n),\ n \geq 1\}$ be a sequence of random variables. For $\delta \in (0,1)$ and s samples, let

$$w = \left\lfloor 8\log(e^{\frac{1}{8}}\delta^{-1}) \wedge \frac{s}{2} \right\rfloor,$$

be the number of blocks, and $m = \lfloor s/w \rfloor$ be the block-length. For $j = 1, 2, \ldots, w$, let

$$\hat{U}_j = \frac{1}{m} \sum_{i=(j-1)m+1}^{jm} U_i,$$

be the sample mean of block j. The median-of-means estimator is computed as follows:

$$\bar{U}_s = \text{median}\{\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_w\}.$$

Definition 5.3 (Median-of-Means Estimator for Reward Rate). Given a renewal reward process $\{(X_n^{(k)}, R_n^{(k)}),\ n \geq 1\}$, for any $b \in \mathcal{B}$, let

$$\begin{aligned} V_i^{(k)} &= R_i^{(k)} \mathbb{I}_{\{X_i^{(k)} \leq b\}}, \\ U_i^{(k)} &= X_i^{(k)} \wedge b, \end{aligned} \tag{16}$$

for all $i = 1, 2, \ldots, s$. Let $\bar{U}_s^{(k)}(b)$ and $\bar{V}_s^{(k)}(b)$ be the median-of-means estimators for $\{(U_i^{(k)}), i \leq s\}$ and$\{(V_i^{(k)}), i \leq s\}$, respectively. Then, the median-of-means estimator $\bar{r}_s^{(k)}(b)$ for the reward rate $r^{(k)}(b)$ is defined as follows:

$$\bar{r}_s^{(k)}(b) = \frac{\bar{V}_s^{(k)}(b)}{\bar{U}_s^{(k)}(b)}, \tag{17}$$

Intuitively, the median-of-means estimator boosts the confidence of a sequence of independent weak estimators (sample mean estimator in this case) by taking the median of them. This successfully eliminates the effect of the outliers due to the heavy tails, and provides fast convergence. In the following, we analyze the performance of the median-of-means estimator for the reward rate.

Proposition 5.4 (Concentration Inequalities for Renewal Processes). *Consider a renewal reward process $\{(X_n^{(k)}, R_n^{(k)}),\ n \geq 1\}$ that satisfies the moment assumption (1) for some $\gamma > 0$. For any $b \in \mathcal{B}$, let $U_i^{(k)} = X_i^{(k)} \wedge b$ and $V_i^{(k)} = R_i^{(k)} \mathbb{I}_{\{X_i^{(k)} \leq b\}}$ for all $i \leq s$, and*

$$\max\left\{\mathbb{E}[|U_1^{(k)} - \mathbb{E}[U_1^{(k)}]|^{1+\gamma}], \mathbb{E}[|V_1^{(k)} - \mathbb{E}[V_1^{(k)}]|^{1+\gamma}]\right\} = u.$$

Then, for any $b \in \mathcal{B}$ and $\delta \in (0,1)$, the median-of-means estimator, $\bar{r}_s(b)$, satisfies the following:

$$\mathbb{P}\Big(\bar{r}_s^{(k)}(b) \leq r^{(k)}(b) - \Delta_0(\epsilon(\delta))\Big) \leq \delta,$$

where

$$\epsilon(\delta) = (12u)^{\frac{1}{1+\gamma}} \left(\frac{16\log(2e^{\frac{1}{8}}\delta^{-1})}{s}\right)^{\frac{\gamma}{1+\gamma}}, \tag{18}$$

and

$$\Delta_0(\epsilon) = \frac{\epsilon}{\mathbb{E}[U_1^{(k)}] + \epsilon}\Big(1 + r^{(k)}(b)\Big).$$

Proposition 5.4 uses the concentration properties of the median-of-means estimator to yield an upper confidence bound on the reward rate in conjunction with Lemma 5.1.

Proof. By taking $p = \infty$, Proposition 5.1 with union bound yields the following upper bound:

$$\mathbb{P}\Big(\bar{r}_s^{(k)}(b) \leq r^{(k)}(b) - \Delta_0(\epsilon)\Big) \leq \mathbb{P}(\bar{U}_s^{(k)} > \mathbb{E}[U_1^{(k)}] + \epsilon) + \mathbb{P}(\bar{V}_s^{(k)} \leq \mathbb{E}[V_1^{(k)}] - \epsilon).$$

Taking $\epsilon = \epsilon(\delta)$, application of Lemma 2 in [8] to each term on the RHS above gives the upper bound. □

Note that the inequalities in Prop. 5.4 require knowledge of arm statistics, which are assumed to be unknown in our case. For the learning problem, we make the following assumptions.

Assumption 2. For the BwI problem, we assume that the following quantities are known a priori:

(1) $\gamma \in (0, 1]$ and $u \in \mathbb{R}$ such that

$$\max\Big\{\mathbb{E}\big[\big|X_1^{(k)} - \mathbb{E}[X_1^{(k)}]\big|^{1+\gamma}\big], \mathbb{E}\big[\big|R_1^{(k)} - \mathbb{E}[R_1^{(k)}]\big|^{1+\gamma}\big]\Big\} \leq u, \tag{19}$$

for all $k \in \mathcal{K}$.

(2) A lower and upper bound on the mean completion time and reward, respectively:

$$\begin{aligned} R_{max} &> \max_{k\in\mathcal{K}} \mathbb{E}[R_1^{(k)}], \\ \mu_{min} &\leq \min_{k\in\mathcal{K}} \mathbb{E}[X_1^{(k)} \wedge b_1]. \end{aligned} \tag{20}$$

The following corollary to Prop. 5.4, which requires much less knowledge about the arm statistics, will be fundamental in algorithm design and analysis.

Corollary 5.5. *Given the parameters in Assumption 2, let*

$$\Delta(x) = \Big(1 + R_{max}/\mu_{min}\Big)\frac{x}{\mu_{min} + x}, \tag{21}$$

for any $x > 0$. Then, the following inequality holds for any $\delta > 0$:

$$\mathbb{P}\Big(\bar{r}_s^{(k)}(b) \leq r^{(k)}(b) - \Delta\big(\epsilon(\delta)\big)\Big) \leq \delta, \tag{22}$$

where $\epsilon(\delta)$ is defined in (18) and $\bar{r}_s^{(k)}(b)$ is the median-of-means estimator for the reward rate.

Proof. With the corresponding parameter choices, for any $\epsilon > 0$, it is easy to show the following:

$$\Delta(\epsilon) > \Delta_0(\epsilon),$$

where Δ_0 is defined in (13). This implies that

$$\mathbb{P}\Big(\bar{r}_s^{(k)}(b) \leq r^{(k)}(b) - \Delta(\epsilon)\Big) \leq \mathbb{P}\Big(\bar{r}_s^{(k)}(b) \leq r^{(k)}(b) - \Delta_0(\epsilon)\Big).$$

Thus, we get the inequality in (22). □

In the next subsection, we examine a specific information structure of the BwI problem.

5.2 Information Structure

The decision $(k, b) \in \mathcal{K} \times \mathcal{B}$ yields the following stochastic observation:

$$Y_n(k, b) = \Big(\mathbb{I}_{\{X_n^{(k)} \leq b\}}, X_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \leq b\}}, R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \leq b\}}\Big).$$

For any pair of interrupt times $b_l < b_{l'}$, we have the following relation between the observation vectors:

$$Y_n(k, b_l) \in \sigma\big(Y_n(k, b_{l'})\big),$$

where $\sigma(Z)$ denotes the sigma field of a random variable Z. Accordingly, the completion time and reward for (k, b_l) is obtained from the observation for $(k, b_{l'})$ as follows:

$$\begin{aligned} X_n^{(k)} \wedge b_l &= \big(X_n^{(k)} \wedge b_{l'}\big) \wedge b_l, \\ R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \leq b_l\}} &= \big(R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \leq b_{l'}\}}\big) \mathbb{I}_{\{X_n^{(k)} \leq b_l\}}. \end{aligned} \tag{23}$$

This immediately implies that any observation vector for the decision $(k, b_{l'})$ can be directly used for (k, b_l) via the transformation in (23). Note that the feedback for the decision (k, b_l) provides some information for $(k, b_{l'})$ for $l < l'$, but it is not very useful as the right tail is censored. Therefore, the information structure is asymmetric.

As a consequence of the aforementioned structure, for fixed k, each decision (k, b_l) has available samples from $(k, b_{l'})$ for all $l' > l$. In order to quantify the improvement in the convergence rate due to the information structure, for $l \in \{1, 2, \ldots, L\}$, let $T_l^{(k)}(n) = \sum_{i=1}^{n} \mathbb{I}_{\{I_i = k, B_i^{(k)} = b_l\}}$ be the number of (k, b_l) decisions among the first n decisions. Then, the effective sample size of the decision (k, b_l) is as follows:

$$\bar{T}_l^{(k)}(n) = \sum_{j \geq l} T_j^{(k)}(n). \tag{24}$$

Note that the effective sample size $\bar{T}_l^{(k)}(n)$ is significantly larger than $T_l^{(k)}(n)$, which implies a much faster convergence in estimation. The effect of this structure in the regret performance will be examined in Section 6.

In the following section, we propose a UCB-type algorithm that exploits the information structure.

5.3 UCB-BwI Algorithm

In this subsection, we will introduce a low-complexity and order-optimal algorithm called the UCB-BwI Algorithm and denoted as $\pi^{\mathtt{BwI}}$.

Design strategy: If all arm statistics, thus $\{\big(r^{(k)}(b)\big),\ (k, b) \in \mathcal{K} \times \mathcal{B}\}$ are known, then one can express the optimal static policy π_n^* as the solution of the following optimization problem:

$$(I_n, B_n^{(I_n)}) = \underset{(k,b) \in \mathcal{K} \times \mathcal{B}}{\arg\max}\ r^{(k)}(b),\ \forall k \in \mathcal{K}, \tag{25}$$

In the absence of the knowledge of $r^{(k)}(b)$, the controller has to learn the arm statistics while maximizing the cumulative reward. The basic idea behind UCB-BwI is to use upper confidence bounds proposed in Corollary 5.5 as a surrogate for $r^{(k)}(b)$.

Observation sequence: Under the UCB-BwI Algorithm, the sequence of observations for (k, b_l) is the following:

$$\{(X_i^{(k)} \wedge b_l, R_i^{(k)} \mathbb{I}_{\{X_i^{(k)} \leq b_l\}}) : 1 \leq i \leq n, \pi_i^{\mathtt{BwI}} = (k, b_{l'}), \forall l \leq l'\}. \tag{26}$$

Each time a decision $\pi_i^{\mathtt{BwI}} = (k, b_{l'})$ is made for $l' \geq l$, a sample is obtained for (k, b_l) via the transformation in (23). Recall that the number of samples for (k, b_l) after n-th decision is $\bar{T}_l^{(k)}(n)$, the effective size defined in (24).

*Definition 5.6 (*UCB-BwI *Algorithm).* For (k, b_l), the median-of-means estimator for the observation sequence (26) of size $s = \bar{T}_l^{(k)}(n)$, denoted by $\bar{r}_{n,s}^{(k)}(b_l)$, is computed by using (17). Let

$$\beta_{u,\gamma} = (12u)^{\frac{1}{1+\gamma}} 32^{\frac{\gamma}{1+\gamma}}, \tag{27}$$

where u is the centralized moment of order $(1 + \gamma)$ defined in (19), and

$$\epsilon_{n,s} = \beta_{u,\gamma} \left[\frac{\log\left(\sqrt{2} e^{\frac{1}{16}} (n+1)^2\right)}{s} \right]^{\frac{\gamma}{1+\gamma}}, \forall n, s \in \mathbb{N}. \tag{28}$$

Then, the arm and interrupt time pair is chosen under UCB-BwI as follows:

$$\left(I_{n+1}, B_{n+1}^{(I_{n+1})}\right) \in \underset{(k\times b_l)\in\mathcal{K}\times\mathcal{B}}{\arg\max} \left\{\bar{r}^{(k)}_{n,\bar{T}_l^{(k)}(n)}(b_l) + \Delta\left(\epsilon_{n,\bar{T}_l^{(k)}(n)}\right)\right\},$$

where Δ is defined in (21). At each epoch, the above optimization can be performed in two steps: first, the optimal interrupt time $B_{n+1}^{(k)}$ is determined for each arm k, and then the optimal $(k, B_{n+1}^{(k)})$ is chosen. The UCB-BwI Algorithm is summarized in Algorithm 1.

Remark 3. We know that the optimal static policy under known statistics selects the arm with the maximum reward rate. A natural idea under unknown statistics is to use an empirical estimator for the reward rate and add an upper-confidence correction to encourage exploration. However, due to the fact that the distributions can be potentially heavy-tailed, the following modifications must be made compared to traditional UCB-type algorithms:

(1) $\bar{r}^{(k)}(b)$ is a median-of-means estimator instead of the usual sample mean estimator. As it was noted in Section 5.1, the empirical reward rate will not have exponential concentration around the reward rate, whereas the median-of-means estimator does.

(2) The UCB correction term typically involves $T_l^{(k)}(n)$, the number of (k, b_l) decisions in the first n epochs. As mentioned in Section 5.2, each decision yields information about others as a result of the information structure. Therefore, the UCB correction term involves the effective sample size, $\bar{T}_l^{(k)}(n)$, of the relevant decisions. Additionally, the heavy-tail parameter γ appears in the exponent of the UCB correction term. Note that if $\gamma = 1$, i.e., $Var(X_1^{(k)})$ and $Var(R_1^{(k)})$ exist, then we obtain the same convergence rate and the correction term as the sub-Gaussian case up to a coefficient, which implies the effectiveness of the estimator. On the other hand, the coefficients $\beta_{u,\gamma}$ can be very large, which makes the UCB conservative in practice.

In order to investigate the value of exploiting the information structure in the numerical examples, we also propose the following UCB-based naive algorithm which does not exploit the information structure.

*Definition 5.7 (*UCB-N *Algorithm).* Let $\bar{r}_{n,T_l^{(k)}(n)}(b)$ be the median-of-means estimator based on observations from only (k, b) decisions up to n-th epoch. The UCB-N Algorithm, which is denoted as π^{N}, makes a decision as follows:

$$(I_{n+1}, B_{n+1}^{(I_{n+1})}) \in \underset{(k,b_l)\in\mathcal{K}\times\mathcal{B}}{\arg\max} \left\{\bar{r}^{(k)}_{n,T_l^{(k)}(n)}(b_l) + \Delta\left(\epsilon^{(k)}_{n,T_l(n)}\right)\right\},$$

where $\epsilon_{n,s}$ is defined in (28).

6 PERFORMANCE ANALYSIS

In this section, we analyze the performance of the UCB-BwI Algorithm by providing a distribution-dependent regret upper bound, and then showing that this performance is order-optimal in K, L and τ by a regret lower bound.

6.1 Regret Upper Bound for UCB-BwI

The main result of this section is the following regret upper bound for the UCB-BwI Algorithm.

Theorem 6.1 (Regret Upper Bound for UCB-BwI). *Under Assumption 2, let (R_{max}, μ_{min}, u) be given as in (19) and (20). Then, the regret under π^{BwI} is upper bounded for all $\tau > 0$ as follows:*

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

Algorithm 1: `UCB-BwI` Algorithm

```
input: $\mathcal{B}$: Set of interrupt times, $\tau$: Time
Initialization:
$S_0^\pi = 0$;
$n = 0$;
$Rew_n = 0$;
for $k = 1, 2, \ldots, K$ do
    for $l = 1, 2, \ldots, L$ do
        $n = n + 1$;
        $I_n = k, B_n^{(k)} = b_l$;
        $S_n^\pi = S_{n-1}^\pi + \min\{X_n^{(k)}, b_l\}$;
        if $X_n^{(I_n)} \le B_n^{(I_n)}$ and $S_n^\pi \le \tau$ then
            $Rew_n = Rew_{n-1} + R_n^{(k)} \cdot \mathbb{I}_{\{X_n^{(k)} \le b_l\}}$;
while $S_n^\pi \le \tau$ do
    $n = n + 1$;
    for $k = 1, 2, \ldots, K$ do
        Compute $B_n^{(k)}$; // Interrupt time for arm $k$.
    Start the process $I_n$; // Arm selection at step $n$. $S_n^\pi = S_{n-1}^\pi + \min\{X_n^{(I_n)}, B_n^{(I_n)}\}$;
    if $X_n^{(I_n)} \le B_n^{(I_n)}$ and $S_n^\pi \le \tau$ then
        $Rew_n = Rew_{n-1} + R_n^{(k)} \cdot \mathbb{I}_{\{X_n^{(k)} \le b_l\}}$;
```

where

$$C^{(k)} = 128(24u)^{\frac{1}{\gamma}} \Big(\frac{1 + \frac{R_{max}}{\mu_{min}}}{\mu_{min}} \Big)^{\frac{\gamma+1}{\gamma}} \Big[\big(\frac{1}{d_{min}^{(k)}}\big)^{\frac{1}{\gamma}} + \big(\frac{1}{d^{(k)}}\big)^{\frac{1}{\gamma}} \Big], \tag{29}$$

for

$$\begin{aligned} d_{min}^{(k)} &= \min_{l \neq l^*(k)} \Big\{ r^{(k)}(b_k^*) - r^{(k)}(b_l) \Big\}, \\ d^{(k)} &= \max_{(k',b)} r^{(k')}(b) - r^{(k)}(b_k^*). \end{aligned} \tag{30}$$

According to Theorem 6.1, the regret grows at a rate $O(K\log(\tau) + KL)$. Similar to the traditional bandit settings, an increasing number of potential actions obliges the controller to make more suboptimal decisions in the learning process, leading to a higher regret. Therefore, a larger set of interrupt times, $\mathcal{B}$, incurs a higher regret as expected. However, as a result of the specific information structure, the coefficient of the time-dependent term in the regret, $C^{(k)}$, is independent of $L = |\mathcal{B}|$.

Proof. We will provide a proof sketch for Theorem 6.1 here. The complete proof can be found in Appendix B. A similar proof strategy is followed in [35] in the context of budgeted bandits.

First, note that the decision times are random and depend on the sample path as well as the policy. Moreover, the total number of decisions, $N_\pi(\tau)$, is a random variable that depends on the realizations. In the following, we tackle these difficulties in two steps: first we express the regret of a policy in terms of the number of suboptimal decisions. In the second step, we analyze the number of suboptimal decisions under $\pi^{\mathtt{BwI}}$ to obtain a regret upper bound.

Step 1. Regret in terms of the number of suboptimal decisions:

The objective in this step is to express the regret in terms of the number of epochs when a suboptimal decision (k, b) is made.

Let $r^* = \max_{k,b}\ r^{(k)}(b)$. The overshoot function, which upper bounds the expected reward of the last initiated and uncompleted task, is defined as follows:

$$\Phi_\pi(\tau) = r^* \sum_{k,l} \mathbb{P}\big(\pi_{\nu_\pi(\tau)} = (k, b_l)\big)\mathbb{E}[X^{(k)}_{\nu_\pi(\tau)} \wedge b_l], \tag{31}$$

where $\nu_\pi(\tau) = N_\pi(\tau) + 1$ is the first passage time under π. Then, for any $\tau > 0$, the regret under π can be upper bounded as:

$$\overline{Reg}_\pi(\tau) \leq \mathbb{E}\Big[\sum_{n=1}^{N_\pi(\tau)} \sum_{k,l} \mathbb{I}_{\{\pi_n=(k,b_l)\}}\Big(r^* - r^{(k)}(b_l)\Big)\mu_{max}\Big] + 2\max_k\ \mathbb{E}[R_1^{(k)}] + \Phi_\pi(\tau). \tag{32}$$

The RHS of (32) can be interpreted as follows: for each decision (k, b_l), the difference $r^* - r^{(k)}(b_l)$ corresponds to the regret rate (per unit time), and multiplication of the regret rate by μ_{max}, an upper bound for the average completion time, yields an upper bound for the regret of an epoch.

Let $\bar{n} > 0$ be a given integer. Under an admissible policy π, the regret upper bound in (32) can be decomposed into two parts:

$$\begin{aligned}\overline{Reg}_\pi(\tau) \leq \sum_{k,l} \mathbb{E}[T_l^{(k)}(\bar{n})]\Big(r^* - r^{(k)}(b_l)\Big)\mu_{max} \\ + KL \cdot r^*\mu_{max} \sum_{n>\bar{n}} \mathbb{P}(N_\pi(\tau) > n) + \Phi_\pi(\tau) + 2\max_k\ \mathbb{E}[R_1^{(k)}].\end{aligned} \tag{33}$$

Intuitively, the upper bound in (33) corresponds to the regret when the process continues for $\bar{n}$ pulls, and then maximal possible regret is incurred for every pull until the time τ expires. Therefore, the natural choice for $\bar{n}$ is a high-probability upper bound for $N_\pi(\tau)$. For any $\delta \in (0, \mu_{min})$, let $\bar{n} = \bar{n}_{\delta,\tau} = \tau/(\mu_{min} - \delta)$. Then, the following result can be obtained:

$$\sum_{n>\bar{n}_{\delta,\tau}} \mathbb{P}\big(N_\pi(\tau) > n\big) = O\Big(\frac{1}{\delta^2}\Big), \tag{34}$$

by using the renewal relation $\{N_\pi(\tau) > n\} \subset \{S_n^\pi \leq \tau\}$ and a concentration inequality for the controlled random walk S_n^π. This implies that:

$$\overline{Reg}_\pi(\tau) \leq \sum_{k,l} \mathbb{E}[T_l^{(k)}(\bar{n}_{\delta,\tau})]\Big(r^* - r^{(k)}(b_l)\Big)\mu_{max} + \Phi_\pi(\tau) + O\Big(\frac{KL}{\delta^2}\Big). \tag{35}$$

The results we obtained so far dealt with the continuous nature of the problem, and provided a connection between the continuous-time process and the number of pulls $\big\{T_l^{(k)}(n)\big\}$ under an admissible policy π. In the last step, we investigate the performance of $\pi^{\mathtt{BwI}}$ in terms of $\big\{T_l^{(k)}(n)\big\}$ to prove the upper bound.

Step 2: Number of suboptimal decisions under $\pi^{\mathtt{BwI}}$:

In this part, we investigate the performance of the UCB-BwI Algorithm. Let $d^{(k)}$ and $d_{min}^{(k)}$ be as defined in (30). Under $\pi = \pi^{\text{BwI}}$, for all k, we have the following results for the number of epochs a suboptimal interrupt time is chosen:

$$\mathbb{E}[T_l^{(k)}(\bar{n})] \leq \left(C_l^{(k)} - C_{l+1}^{(k)}\right)^+ \log(\bar{n}) + O(1),\ \forall l > l^*(k),$$
$$\mathbb{E}[T_l^{(k)}(\bar{n})] = O\Big(\big(d_{min}^{(k)}\big)^{-\frac{1+\gamma}{\gamma}}\Big),\ \forall l < l^*(k).$$

where $x^+ = \max\{x, 0\}$ for any $x \in \mathbb{R}$, and

$$C_l^{(k)} = 128(24u)^{\frac{1}{\gamma}}\Big(\frac{1 + r^{(k)}(b_l)}{d_l^{(k)}\mathbb{E}[X_1^{(k)} \wedge b_l]}\Big)^{\frac{\gamma+1}{\gamma}}.$$

From this, the following upper bound for the expected number of suboptimal time decisions can be established:

$$\sum_{l \neq l^*(k)} \mathbb{E}[T_l^{(k)}(\bar{n})] \leq \max_{l > l^*(k)} C_l^{(k)} \log(\bar{n}) + O\big(\frac{L}{(d_{min}^{(k)})^\gamma}\big).$$

Consequently, the coefficient of the logarithmic term is independent of the size of the interrupt set, which implies low regret even for a large interrupt set. Also, this result implies that the overshoot under π^{BwI} is $O(1)$.

For each suboptimal arm k, we can prove the following upper bound for the expected number of epochs a suboptimal arm is chosen together with its optimal interrupt time:

$$\mathbb{E}[T_{l^*(k)}^{(k)}(\bar{n})] \leq 128(24u)^{\frac{1}{\gamma}}\Big(\frac{1 + \frac{R_{max}}{\mu_{min}}}{d^{(k)}\mu_{min}}\Big)^{\frac{\gamma+1}{\gamma}} \log(\bar{n}) + O(1). \tag{36}$$

The equation (36) implies that each suboptimal arm, paired with its optimal interrupt time, is chosen for at most logarithmically many epochs under the UCB-BwI Algorithm.

Substituting the upper bounds for $\mathbb{E}[T_l^{(k)}(\bar{n})]$ to (35), and minimizing the resulting upper bound over $\delta \in (0, \mu_{min})$ yields the the regret upper bound in Theorem 6.1. □

6.2 Regret Lower Bound for Admissible Policies

In this section, we will analyze the regret lower bounds for the class of admissible policies that include UCB-BwI. As it will be seen, the regret under any such policy grows at a rate $\Omega(K \log \tau)$, which implies that UCB-BwI is order optimal.

Consider a K-armed bandit with $\mathcal{B} = \{b_1, b_2, \dots, b_L\}$. For each arm $k \in \mathcal{K}$, the arm process is distributed according to a parametric distribution $(X_n^{(k)}, R_n^{(k)}) \sim P_{\theta^{(k)}}$ where $\theta^{(k)} \in \Theta_k$ for a parameter set Θ_k, and let $\Theta = \Theta_1 \times \Theta_2 \times \ldots \times \Theta_K$ be the parameter set for the problem. Note that the arm distributions do not need to belong to the same family of distributions in this case. As in (2), an observation from arm k with interrupt time b_l is denoted as a vector:

$$Y_n(k, b) = \Big(\mathbb{I}_{\{X_n^{(k)} \leq b\}},\ X_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \leq b\}},\ R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \leq b\}}\Big).$$

Given $\theta = (\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(K)})$, the distribution of $Y_n(k, b_k^*)$ is denoted as $P_\theta^{(k)}$. For each $k \in \mathcal{K}$, we define the following subset of Θ:

$$\Theta_k^* = \{(\theta^{(1)}, \dots, \theta^{(K)}) \in \Theta : r^{(k)}(b_k^*) \geq r^{(k')}(b_{k'}^*),\ \forall k' \in \mathcal{K}\}.$$

Note that Θ_k^* is the set of parameters for which arm k is optimal. After these definitions, we make the following assumptions which are analogous to those in [25].

Assumption 3. For a given $\mathcal{B}$ and Θ, we assume the following:

- Interior of any Θ_k^* is non-empty, i.e., $int(\Theta_k^*) \neq \emptyset$ for any k.
- The true parameter $\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)})$ lies in the interior of a partition:
$$\theta \in int(\Theta_1^*),$$
where we assumed that Arm 1 is optimal without loss of generality.

If this technical assumption is satisfied by the problem instance, the following regret lower bound holds for all admissible policies.

Theorem 6.2 (Regret Lower Bound for Admissible Policies). *Under Assumption 3, consider an admissible policy π that satisfies the uniform optimality condition:*

$$\mathbb{E}[T_l^{(k)}(n)] = o(n^\alpha),\ \forall \alpha > 0, n \longrightarrow \infty,\ \forall k. \tag{37}$$

Let

$$D_k(\theta||\theta') = \mathbb{E}[\log \frac{dP_\theta^{(k)}}{dP_{\theta'}^{(k)}}(Y_1(k, b_k^*))],$$

be the Kullback-Leibler (KL) divergence between $P_\theta^{(k)}$ and $P_{\theta'}^{(k)}$. Then, the following lower bound holds for any $\delta > 0$:

$$\liminf_{\tau \to \infty} \frac{\overline{Reg}_\pi(\tau)}{\log\left(\frac{\tau}{\mu_{max}+\delta}\right)} \geq \mu_{min} \sum_{k \neq 1} \frac{d^{(k)}}{\inf\limits_{\theta' \in \Theta_k^*} D_k(\theta||\theta')} - O\Big(\frac{1}{\delta^{1+\gamma}}\Big), \tag{38}$$

where $d^{(k)} = \max\limits_{(k,b)} r^{(k)}(b) - r^{(k)}(b_k^*)$, $\mu_{min} = \min\limits_{(k,b)} \mathbb{E}[X \wedge b]$ and $\mu_{max} = \max\limits_{k} \mathbb{E}[X_1^{(k)}]$.

Theorem 6.2 implies that the regret under a "good" policy that satisfies (37) is $\Omega(K \log(\tau))$ for a K-armed bandit problem with a set of L interrupt times. Recall that the regret under UCB-BwI is $O(K \log(\tau))$ by Theorem 6.1. Hence, these two results imply that UCB-BwI is order-optimal in K, L and τ.

Proof. The proof is given in Appendix C. □

6.3 Discussion

It is interesting to note that the regret under UCB-BwI grows over time at a rate $O(K \log \tau)$, which is independent of L. The main reason for this is the information structure of the problem. If instead there were no such correlation between the interrupt time decisions, it is straightforward to obtain a regret lower bound of $\Omega(KL \log \tau)$ by using a Lai-Robbins style approach. As such, a scaling gain of $O(1/L)$ is achieved from the use of the information structure by our design. This result is similar in spirit to earlier results in [13] or [12, 17] in that they also exploit different information structures to achieve scaling gains. However, our setting has a different particular information structure that stems from the dynamics of the renewal processes, which is optimally exploited in terms of τ, K and L by UCB-BwI.

7 NUMERICAL RESULTS

To corroborate the theoretical results we obtained in Section 6, we investigate the regret performance of UCB-BwI in various settings. In order to investigate the effect of exploiting the information structure, we also evaluate the performance of UCB-N.

(1) Adaptive Task Scheduling with Interruptions:

In the first example, we consider the adaptive task scheduling problem during the busy period of a single server. The completion time of n-th task is $X_n^{(1)}$, learned by the controller via feedback only after the completion. At the beginning of the task, the controller determines an interrupt time $B_n^{(1)}$. If the task is not completed within $B_n^{(1)}$ time units after the initiation, it is discarded and the succeeding task is initiated. The goal is to maximize the number of completed tasks within $[0, \tau]$ by learning the optimal interrupt time, thus $R_n^{(1)} = 1$ for all n.

In communication systems with ARQ control, the task completion times have a heavy-tailed distribution [21, 32]. In order to model such systems, we consider a scenario where the task completion time has a $Pareto(1, 1.4)$ distribution. We consider the following set of interrupt times:

$$\mathcal{B}_L = \{3, 6, 9, \ldots, 3(L-1), \infty\},\ L > 1, \tag{39}$$

In Figure 3, the regret performances of `UCB-BwI` and `UCB-N` are presented for $L \in \{4, 12\}$.

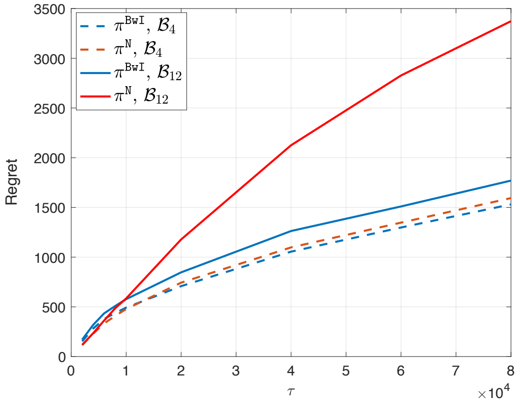

Fig. 3. The regret performances of π^{BwI} and π^{N} for the adaptive task scheduling problem. An expanding set of interrupt times has a drastic effect on the regret under π^{N} while the increase in the regret under π^{BwI} is not significant.

From Figure 3, we observe that the regret under `UCB-N` grows significantly as the size of the information set increases from 4 to 12. On the other hand, the increase in the regret under `UCB-BwI` is considerably small. This verifies the result of Theorem 6.1: the exploitation of the regret yields significant scaling gains over time as the interrupt set expands.

(2) Adaptive Task Scheduling with Multiple Types:

In the second example, we investigate the performance of `UCB-BwI` in a bandit setting with three arms of distinct statistical characteristics. This example models the adaptive task scheduling problem with three task types. For a given time horizon τ, the objective of the scheduler is to maximize the number of completed tasks in $[0, \tau]$ by learning the optimal arm and interrupt time pair. In order to exhibit the effectiveness of the non-parametric approach, we consider a highly diverse set of completion time distributions:

- Arm 1: $X_n^{(1)} \sim Pareto(1, 1.4)$.
- Arm 2: $X_n^{(3)} \sim Exp(1/3)$.
- Arm 3: $X_n^{(2)} \sim Uniform(0, 6)$.

The reward rates for these distributions are plotted in Figure 4. From Figure 4, we observe that both Arm 2 and Arm 3 yield higher reward rates than Arm 1 without interruption. However, Arm 1 yields the optimal reward rate if it is paired with the optimal interruption time. Therefore, an

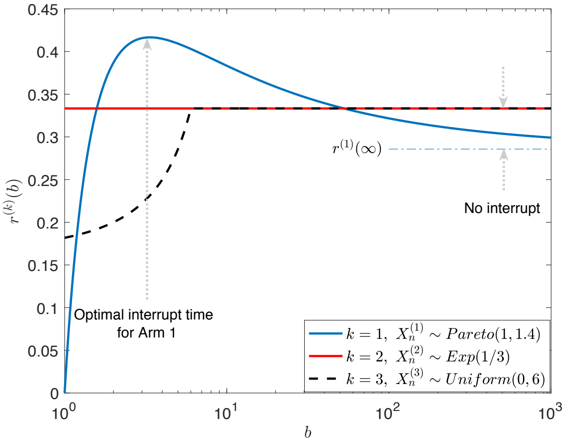

Fig. 4. The renewal reward rates for $Pareto(1, 1.4)$, $Exp(1/3)$ and $Uniform(0, 6)$ completion time distributions. Without interruption, Arm 2 and Arm 3 yield higher reward rates than Arm 1. However, Arm 1 achieves the optimal reward rate with optimal interruption.

algorithm must learn not only the first-order statistics but the complete distribution of an arm to achieve sublinear regret.

The regret performance of the UCB-BwI Algorithm is presented in Figure 5 for the set of interrupt times $\mathcal{B}_L$ defined in (39). In this example, large interrupt sets are considered to include potentially large optimal interrupt times due to the unknown and diverse arm statistics. From Figure 5, we

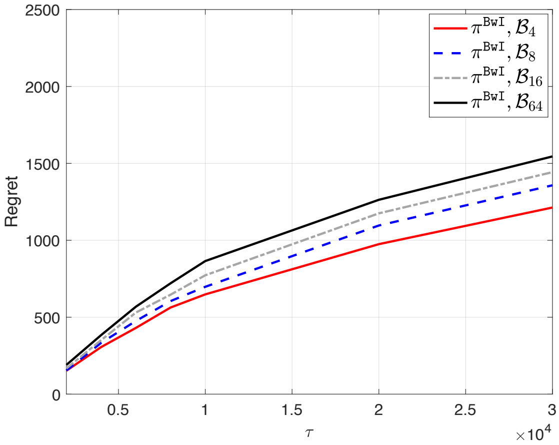

Fig. 5. The regret performance of UCB-BwI for a three-armed bandit with $Pareto(1, 1.4)$, $Exp(1/3)$ and $Uniform(0, 6)$ completion time distributions.

observe that although the set of interrupt times grows significantly from 4 to 64 possible interrupt times, this has little effect on the regret performance of the UCB-BwI. This suggests that BwI provides an effective solution to the statistically diverse problem instances where the set of interrupt times is large.

In Figure 6, for each $(k, b) \in \mathcal{K} \times \mathcal{B}_{16}$, we present the fraction of (k, b) decisions for a time horizon $\tau = 4 \times 10^4$ under the UCB-BwI Algorithm. From Figure 6, we observe that b_L is chosen more frequently than other suboptimal interrupt times under UCB-BwI. Note that b_L is the most informative decision among all interrupt times since it provides a complete feedback for each decision as a result of the information structure. Consequently, UCB-BwI makes a balanced exploration of b_L, and the optimal interrupt time for each arm is learned with low regret even for a large set of interrupt times.

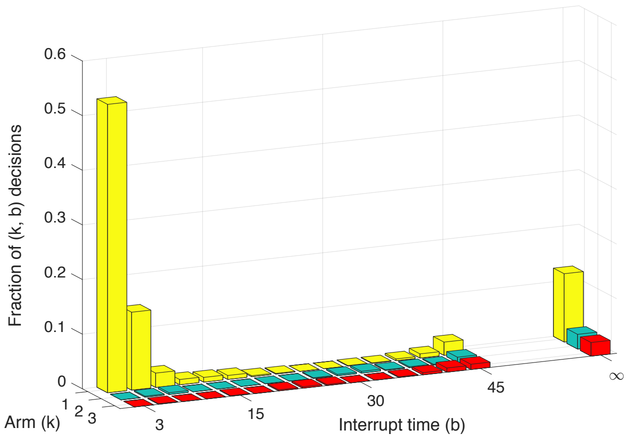

Fig. 6. The fraction of (k, b) decisions under the UCB-BwI Algorithm. UCB-BwI makes a balanced exploration of the most informative interrupt time b_L yielding low regret.

8 CONCLUSIONS

In this work, we introduced a novel continuous-time multi-armed bandit framework where each arm corresponds to a distinct renewal process. In this setting, each arm pull initiates a task, and a reward is obtained after the completion of the task. We showed that enabling the controller to interrupt a task yields significant gains in the expected cumulative reward within a given time interval. We characterized the optimal policy given arm statistics, and observed that all heavy-tailed and some light-tailed completion time distributions require such an interrupt mechanism for optimal performance. For the learning problem, the interrupt mechanism obliges the learner to learn the whole distribution rather than just the mean, and this is done via censored observations due to task interruption. For this purpose, we proposed a non-parametric algorithm based on median-of-means estimator called the UCB-BwI Algorithm. By construction, UCB-BwI exploits the specific information structure of the problem. We proved that the regret under UCB-BwI is $O(K \log \tau + KL)$ for a set of L interrupt times. By a regret lower bound, we also proved that UCB-BwI is order optimal in τ, K and L.

ACKNOWLEDGMENTS

This research is supported in part by the NSF grants: CNS-NeTS-1514260, CNS-NeTS-1717045, CMMI-SMOR-1562065, CNS-ICN-WEN-1719371, CNS-SpecEES-1824337, NeTS 1718203, CPS ECCS 1739189, CMMI 1562276, and ECCS 1609370; ARO grant W911NF-16-1-0259; and the DTRA grants: HDTRA1-15-1-0003, HDTRA1-18-1-0050.

A PROOF OF PROPOSITION 4.6

Proof. The proof consists of two stages: for a given $\tau > 0$, we first find a lower bound for the expected reward under the optimal static policy π^*. Then, we find an upper bound for the maximum expected reward among all admissible policies. The difference between these quantities gives an upper bound for the optimality gap.

(1) Lower bound for $\mathbb{E}[Rew_{\pi^*}(\tau)]$

For any k, let the static policy $\pi^{(k)}$ be defined as $\pi_n^{(k)} = (k, b_k^*)$ for all n where $b_k^* = \arg\max_{b \in \mathcal{B}} r^{(k)}(b)$ is the optimal interrupt time for arm k. Recall that the first hitting time $\nu_{\pi^{(k)}}(\tau) = N_{\pi^{(k)}}(\tau) + 1$ is a stopping time. Thus, by Wald's equation, we have:

$$\mathbb{E}[Rew_{\pi^{(k)}}(\tau)] = \mathbb{E}[\nu_{\pi^{(k)}}(\tau)]\mathbb{E}[R_1^{(k)}\mathbb{I}_{\{X_1^{(k)} \le b_k^*\}}] - \mathbb{E}[R_{\nu^{(k)}(\tau)}^{(k)}\mathbb{I}_{\{X_{\nu^{(k)}(\tau)}^{(k)} \le b_k^*\}}]. \tag{40}$$

By the key renewal theorem, we have $\mathbb{E}[\nu_{\pi^{(k)}}(\tau)] \ge \tau/\mathbb{E}[X_1^{(k)} \wedge b_k^*]$ [2]. Thus, we can lower bound the expected reward under $\pi^{(k)}$ as follows:

$$\mathbb{E}[Rew_{\pi^{(k)}}(\tau)] \ge \tau r^{(k)}(b_k^*) - \mathbb{E}[R_1^{(k)}], \tag{41}$$

since $\mathbb{E}[R_{\nu^{(k)}(\tau)}^{(k)}\mathbb{I}_{\{X_{\nu^{(k)}(\tau)}^{(k)} \le b_k^*\}}] \ge \mathbb{E}[R_{\nu^{(k)}(\tau)}^{(k)}]$ holds, and $R_1^{(k)}$ and $\nu^{(k)}(\tau)$ are independent. Therefore, we have the following lower bound for the expected reward under π^*:

$$\mathbb{E}[Rew_{\pi^*}(\tau)] \ge \tau \max_{k,b} \ r^{(k)}(b) - \max_k \mathbb{E}[R_1^{(k)}], \tag{42}$$

for all $\tau > 0$.

(2) Upper bound for $\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)]$

Consider an admissible policy π. The total time spent until the completion of the n-th task, S_n^π, defined in (3) is a controlled random walk with positive increments, therefore we have $\mathbb{I}_{\{S_n^\pi \le \tau\}} \le \mathbb{I}_{\{S_{n-1}^\pi \le \tau\}}$ with probability 1. Thus, we can upper bound the cumulative reward under π as follows:

$$\begin{aligned}
\mathbb{E}[Rew_\pi(\tau)] &= \mathbb{E}\Big[\sum_{n=1}^{\infty}\sum_{(k,b)} \mathbb{I}_{\{S_n^\pi \le \tau\}}\mathbb{I}_{\{\pi_n=(k,b)\}}R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \le b\}}\Big], \\
&\le \mathbb{E}\Big[\sum_{n=1}^{\infty}\sum_{k,b} \mathbb{I}_{\{S_{n-1}^\pi \le \tau\}}\mathbb{I}_{\{\pi_n=(k,b)\}}\mathbb{E}[R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \le b\}}]\Big], \\
&\le \mathbb{E}\Big[\sum_{n=1}^{N_\pi(\tau)}\sum_{k,b} \mathbb{I}_{\{\pi_n=(k,b)\}}\mathbb{E}[R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \le b\}}]\Big] + \max_k \ \mathbb{E}[R_1^{(k)}],
\end{aligned}$$

where the second line holds as a result of the independence of $R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \le b\}}$ and $\mathbb{I}_{\{S_{n-1}^\pi \le \tau\}}\mathbb{I}_{\{\pi_n=(k,b)\}}$ under an admissible policy. Note that the upper bound above corresponds to the expected reward under π including the reward of the incomplete final task. Since $\mathbb{E}[R_n^{(k)}\mathbb{I}_{\{X_n^{(k)} \le b\}}] = r^{(k)}(b)\mathbb{E}[X_n^{(k)} \wedge b]$, we have the following upper bound:

$$\mathbb{E}[Rew_\pi(\tau)] \le \max_{k,b} r^{(k)}(b) \cdot \mathbb{E}\Big[\sum_{n=1}^{N_\pi(\tau)}\sum_{k,b} \mathbb{I}_{\{\pi_n=(k,b)\}}\big(X_n^{(k)} \wedge b\big)\Big] + \max_k \ \mathbb{E}[R_1^{(k)}], \tag{43}$$

For the first term on the RHS of (43), we have the following inequality:

$$\sum_{n=1}^{N_\pi(\tau)} \sum_{k,b} \mathbb{I}_{\{\pi_n=(k,b)\}} \big(X_n^{(k)} \wedge b\big) \leq \tau, \tag{44}$$

which holds for all sample paths since it is the total time passed before the last activated (but uncompleted) task. Hence, from (43) and (44), we have the following upper bound that holds for any admissible policy π:

$$\mathbb{E}[Rew_\pi(\tau)] \leq \tau \max_{k,b}\ r^{(k)}(b) + \max_k\ \mathbb{E}[R_1^{(k)}], \tag{45}$$

for all $\tau > 0$. Since the above upper bound holds for any π, it also yields an upper bound for $\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)]$.

From (41) and (45), we have the following upper bound for the optimality gap:

$$\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)] - \mathbb{E}[Rew_\pi(\tau)] \leq 2 \cdot \max_k\ \mathbb{E}[R_1^{(k)}], \tag{46}$$

for any $\tau > 0$.

The following holds for the asymptotic performance of π^*:

$$\begin{aligned} 1 \leq \lim_{\tau\to\infty} & \frac{\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)]}{\mathbb{E}[Rew_{\pi^*}(\tau)]} \\ \leq \lim_{\tau\to\infty} & \frac{\tau \max\limits_{k,b}\ r^{(k)}(b) + \max_k \mathbb{E}[R_1^{(k)}]}{\tau \max\limits_{k,b}\ r^{(k)}(b) - \max_k \mathbb{E}[R_1^{(k)}]} = 1, \end{aligned} \tag{47}$$

where the second line follows from (41) and (45). (47) directly implies that

$$\lim_{\tau\to\infty} \frac{\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)]}{\mathbb{E}[Rew_{\pi^*}(\tau)]} = 1,$$

and therefore π^* is asymptotically optimal as $\tau \longrightarrow \infty$. □

B PROOF OF THEOREM 6.1

The proof of Theorem 6.1 builds on three lemmas that we prove in this section. First, we establish the connection between the regret and the number of suboptimal decisions by the following lemma.

Lemma B.1. *For $\mu_{max} = \max_k\ \mathbb{E}[X_1^{(k)}]$ and $r^* = \max_{k,b}\ r^{(k)}(b)$, let*

$$\Phi_\pi(\tau) = r^* \sum_{k,l} \mathbb{P}\big(\pi_{\nu_\pi(\tau)} = (k, b_l)\big)\mathbb{E}[X_{\nu_\pi(\tau)}^{(k)} \wedge b_l].$$

Then, the regret under any admissible policy π can be upper bounded as follows:

$$\overline{Reg}_\pi(\tau) \leq \mathbb{E}\Big[\sum_{n=1}^{N_\pi(\tau)} \sum_{k,l} \mathbb{I}_{\{\pi_n=(k,b_l)\}} \Big(r^* - r^{(k)}(b_l)\Big)\mu_{max}\Big] + 2\max_k\ \mathbb{E}[R_1^{(k)}] + \Phi_\pi(\tau), \tag{48}$$

for all $\tau > 0$.

Proof. In order to prove the lemma, we first find an upper bound for the cumulative reward under the optimal policy $\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)]$, and then find a lower bound for $\mathbb{E}[Rew_\pi(\tau)]$, both in terms of the counting process $N_\pi(\tau)$.

(1) Upper Bound for $\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)]$

By (45), we have the following inequality:

$$\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)] \le r^*\tau + \max_k \mathbb{E}[R_1^{(k)}]. \tag{49}$$

Under policy π, the following holds for any τ by the definition of $N_\pi(\tau)$:

$$\tau \le \sum_{n=1}^{N_\pi(\tau)+1} \sum_{k,l} \mathbb{I}_{\{I_n=k, B_n^{(k)}=l\}}(X_n^{(k)} \wedge b_l). \tag{50}$$

Hence, by (49) and (50), we obtain the following upper bound for the optimal reward:

$$\mathbb{E}[Rew_{\pi^{\text{opt}}}(\tau)] \le \mathbb{E}\Big[\sum_{n=1}^{N_\pi(\tau)} \sum_{k,l} \mathbb{I}_{\{\pi_n=(k,b_l)\}} r^* \mathbb{E}[X_n^{(k)} \wedge b_l]\Big] + \max_k \ \mathbb{E}[R_1^{(k)}] + \Phi_\pi(\tau), \tag{51}$$

(2) Lower Bound for $\mathbb{E}[Rew_\pi(\tau)]$

In order to find a lower bound for $\mathbb{E}[Rew_\pi(\tau)]$, we follow a similar proof technique with the proof of Proposition 4.6.

$$\begin{aligned}\mathbb{E}[Rew_\pi(\tau)] &= \mathbb{E}\Big[\sum_{n=1}^{\infty} \sum_{(k,b)} \mathbb{I}_{\{S_n^\pi \le \tau\}} \mathbb{I}_{\{\pi_n=(k,b)\}} R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}\Big], \\ &\ge \mathbb{E}\Big[\sum_{n=1}^{\infty} \sum_{(k,b)} \mathbb{I}_{\{S_{n-1}^\pi \le \tau\}} \mathbb{I}_{\{\pi_n=(k,b)\}} \mathbb{E}[R_n^{(k)} \mathbb{I}_{\{X_n^{(k)} \le b\}}]\Big], \\ &\ge \mathbb{E}\Big[\sum_{n=1}^{N_\pi(\tau)} \sum_{(k,b)} \mathbb{I}_{\{\pi_n=(k,b)\}} r^{(k)}(b) \mathbb{E}[X_n^{(k)} \wedge b]\Big] - \max_k \ \mathbb{E}[R_1^{(k)}],\end{aligned} \tag{52}$$

The upper bound is obtained by taking the difference between (51) and (52).

□

The number of arm pulls, $N_\pi(\tau)$, in Lemma B.1 is a random variable that depends on the observations. Note that the regret is an increasing function of $N_\pi(\tau)$, therefore we can simplify the analysis by using a high-probability upper bound for $N_\pi(\tau)$.

Lemma B.2. *For $\delta \in (0, \mu_{min})$, let*

$$\bar{n}_{\delta,\tau} = \frac{\tau}{\mu_{min} - \delta},$$

and

$$\chi_\delta(\tau) = \frac{\exp(-2\bar{n}_{\delta,\tau}\delta^2/b_1^2)}{1 - \exp(2\delta^2/b_1^2)}.$$

Then,

$$\overline{Reg}_\pi(\tau) \le \sum_{(k,l):r^{(k)}(b_l)<r^*} \mathbb{E}\big[T_l^{(k)}(\bar{n}_{\delta,\tau})\big]\Big(r^* - r^{(k)}(b_l)\Big)\mu_{max} + \Phi_\pi(\tau) + 2\max_k \ \mathbb{E}[R_1^{(k)}] \\ + K \cdot L \cdot r^* \cdot \chi_\delta(\tau) \cdot \mu_{max}. \tag{53}$$

Proof. By using the fact that $\mathbb{I}_{\{\pi_n=(k,b),S_n^\pi\leq\tau\}} \leq \mathbb{I}_{\{\pi_n=(k,b)\}}$, we can decompose the regret upper bound in (52) as follows:

$$\overline{Reg}_\pi(\tau) \leq \sum_{k,l} \mathbb{E}[T_l^{(k)}(\bar{n})]\Big(r^* - r^{(k)}(b_l)\Big)\mu_{max} + \Phi_\pi(\tau) + 2\max_k \mathbb{E}[R_1^{(k)}] \\ + K\cdot L\cdot r^*\cdot\mu_{max}\cdot\sum_{n>\bar{n}}\mathbb{P}(N_\pi(\tau)>n). \quad (54)$$

for any $\bar{n}\in\mathbb{N}$. By using the renewal relation

$$\{N_\pi(\tau)>n\}\subset\{S_n^\pi\leq\tau\},$$

and since $b_1 \leq b_l$ for all l, we have the following:

$$\{N_\pi(\tau)>n\}\subset\{\sum_{j=1}^n\sum_{k,l}\mathbb{I}_{\{I_j=k,B_j^{(k)}=b_l\}}(X_j^{(k)}\wedge b_l)\leq\tau\} \\ \subset\{\sum_{j=1}^n\sum_{k}\mathbb{I}_{\{I_j=k\}}(X_j^{(k)}\wedge b_1)\leq\tau\}. \quad (55)$$

For $\delta\in(0,\mu_{min})$, let $\bar{n}=\bar{n}_{\delta,\tau}=\tau/(\mu_{min}-\delta)$ in (54). Then, since all random variables in (55) is bounded in $[0,b_1]$, we have the following inequality:

$$\mathbb{P}\big(N_\pi(\tau)>n\big)\leq\mathbb{P}\Big(\sum_{j=1}^n\sum_k\mathbb{I}_{\{I_j=k\}}(X_j^{(k)}\wedge b_1)\leq\tau\Big), \\ \leq\exp\Big(-\frac{2n\delta^2}{b_1^2}\Big),\ \forall n>\bar{n}_{\delta,\tau},$$

by Azuma-Hoeffding inequality. From this, it immediately follows that

$$\sum_{n>\bar{n}_{\delta,\tau}}\mathbb{P}\big(N_\pi(\tau)>n\big)\leq\chi_\delta(\tau)=O\big(1/\delta^2\big),$$

which concludes the proof. □

The results so far made a plausible connection between the regret and the number of suboptimal decisions $T_l^{(k)}(\bar{n}_{\delta,\tau})$. In the final step, we find upper bounds for the number of suboptimal decisions under $\pi^{\mathtt{BwI}}$.

Lemma B.3. *Let $d_l^{(k)}=r^{(k)}(b_k^*)-r^{(k)}(b_l)$, $d^{(k)}=r^*-r^{(k)}(b_k^*)$, and recall that*

$$\max\Big\{\mathbb{E}\big[|X_1^{(k)}-\mathbb{E}[X_1^{(k)}]|^{1+\gamma}\big],\mathbb{E}\big[|R_1^{(k)}-\mathbb{E}[R_1^{(k)}]|^{1+\gamma}\big]\Big\}\leq u.$$

Then we have the following results.

(a) For $l\neq l^*(k)$, let

$$C_l^{(k)}=128(24u)^{\frac{1}{\gamma}}\Big(\frac{1+r^{(k)}(b_l)}{d_l^{(k)}\mathbb{E}[X_1^{(k)}\wedge b_l]}\Big)^{\frac{\gamma+1}{\gamma}},$$

with $C_{L+1}^{(k)} = 0$. *Then,*

$$\sum_{l > l^*(k)} \mathbb{E}[T_l^{(k)}(\bar{n})] \leq \max_{l > l^*(k)} C_l^{(k)} \log\left(1.25\bar{n}\right) + L\frac{\pi^2}{3}, \tag{56}$$

$$\sum_{l < l^*(k)} \mathbb{E}[T_l^{(k)}(\bar{n})] = O\Big(\frac{L}{(d_l^{(k)})^{1+1/\gamma}}\Big). \tag{57}$$

for all $\bar{n} > 0$.

(b) For k *such that* $r^{(k)}(b_k^*) < r^*$, *let*

$$\tilde{C}^{(k)} = 128(24u)^{\frac{1}{\gamma}}\Big(\frac{1 + r^{(k)}(b_k^*)}{d^{(k)}\mathbb{E}[X_1^{(k)} \wedge b_k^*]}\Big)^{\frac{\gamma+1}{\gamma}}.$$

Then,

$$\mathbb{E}[T_{l^*(k)}^{(k)}(\bar{n})] \leq \tilde{C}^{(k)} \log\left(1.25\bar{n}\right) + \frac{\pi^2}{3}. \tag{58}$$

for any $\bar{n}$.

Proof. (a) Let $\Delta_{l,n}^{(k)} = \Delta\Big(\beta_{u,\gamma}\Big[\frac{\log(\sqrt{2}e^{\frac{1}{16}}(n+1)^2)}{\bar{T}_l^{(k)}(n)}\Big]^{\frac{\gamma}{1+\gamma}}\Big)$ from Definition 5.6, and define the following events:

$$E_{1,n}^{(k)} = \Big\{\bar{r}_{n,\bar{T}_{l^*(k)}^{(k)}(n)}^{(k)} + \Delta_{l^*(k),n}^{(k)} \leq r^{(k)}(b_k^*)\Big\},$$

$$E_{2,n}^{(k)} = \bigcup_{l > l^*(k)} \Big\{\bar{r}_{n,\bar{T}_l^{(k)}(n)}^{(k)} > \Delta_{l,n}^{(k)} + r^{(k)}(b_l)\Big\},$$

$$E_{3,n}^{(k)} = \bigcup_{l > l^*(k)} \Big\{\bar{T}_l^{(k)}(n) \leq \frac{C_l^{(k)}}{2} \log(\sqrt{2}e^{\frac{1}{16}}\bar{n}^2)\Big\}.$$

Then, it can be shown by contradiction (similar to Theorem 2.1 in [7]) that the following relation is true:

$$\big\{I_{n+1} = k, B_{n+1}^{(k)} > l^*(k)\big\} \subset E_{1,n}^{(k)} \cup E_{2,n}^{(k)} \cup E_{3,n}^{(k)}. \tag{59}$$

Since $\bar{T}_l^{(k)}(\bar{n}) = \sum_{j \geq l} T_j^{(k)}(\bar{n})$ for all l due to the information structure, we can express $E_{3,n}^{(k)}$ in a more explicit form:

$$E_{3,n}^{(k)} \subset \bigcup_{l > l^*(k)} \Big\{T_l^{(k)}(n) \leq \Big(C_l^{(k)} - C_{l+1}^{(k)}\Big)^+ \log\left(1.25\bar{n}\right)\Big\},$$

where $x^+ = \max\{x, 0\}$. This directly demonstrates the effect of the information structure: the exploration of the suboptimal interrupt time decisions is greatly reduced by the mutual feedback coming from the higher interrupt time decisions.

Thus, we have the following:

$$\sum_{l > l^*(k)} \mathbb{E}[T_l^{(k)}(\bar{n})] \leq \sum_{l > l^*(k)} \Big(C_l^{(k)} - C_{l+1}^{(k)}\Big)^+ \log(1.25\bar{n}) + \sum_{n=1}^{\infty}\Big(\mathbb{P}(E_{1,n}^{(k)}) + \mathbb{P}(E_{2,n}^{(k)})\Big). \tag{60}$$

Note that $\sum_{l > l^*(k)}(C_l^{(k)} - C_{l+1}^{(k)})^+ = \max_{l > l^*(k)} C_l^{(k)}$ with the definition $C_{L+1}^{(k)} = 0$, and the RHS of (60) is upper bounded by $L\pi^2/3$ by the concentration inequality in Lemma 5.5. This concludes the proof of (56).

The equation (57) is proved by using the same argument for $l < l^*(k)$. Note that every sample from the optimal interrupt time is useful for $l < l^*(k)$ due to the information structure. This implies that the effective sample size $\bar{T}_l^{(k)}(n)$ grows linearly over time for all $l < l^*(k)$, i.e., $\bar{T}_l^{(k)}(n) = \Theta(n)$, while only $O(\log \bar{n})$ exploration is necessary. Hence, $\mathbb{E}[T_l^{(k)}(\bar{n})] = O(1)$ for all $l < l^*(k)$.

(b) This part is proved in an identical way as Part (a).

□

The function $\Phi_{\pi^{\texttt{BwI}}}(\tau)$ is $O(1)$ for all light-tailed completion time distributions as the mean of the stopping summand $\mathbb{E}[X_{\nu_\pi(\tau)}^{(k)}]$ is bounded if the variance exists [2]. When the variance of the completion time does not exist, i.e., the completion time distribution is heavy-tailed, not interrupting the task is suboptimal by Corollary 4.4. Therefore, the probability of not interrupting the final task vanishes as $\tau \longrightarrow \infty$, which implies that the overshoot function is bounded.

So far, we established a connection between the regret and the number of suboptimal decisions under a policy, and then provided upper bounds for the number of suboptimal decisions under UCB-BwI. In the final stage, we substitute the upper bounds for $\mathbb{E}[T_l^{(k)}(\bar{n}_{\delta,\tau})]$ into the regret rate expression in (53) to establish the main result:

$$\overline{Reg}_{\pi^{\texttt{BwI}}}(\tau) \leq \sum_{k=1}^{K} \left[C^{(k)} \log\left(\frac{\tau}{\mu_{min} - \delta}\right) + O\Big(\frac{L}{\big(d_{min}^{(k)}\big)^{\frac{1}{\gamma}}}\Big) \right] + O\Big(\frac{KL}{\delta^2}\Big),$$

where $C^{(k)} = \max_{l > l^*(k)} C_l^{(k)} + \tilde{C}^{(k)}$.

C PROOF OF THEOREM 6.2

In order to prove the theorem, we first find a regret lower bound based on the regret rate notion that was used in the proof of Theorem 6.1.

Lemma C.1. *For any $k \in \mathcal{K}$, let*

$$d^{(k)} = \max_{(k', b)} r^{(k')}(b) - r^{(k)}(b_k^*),$$

and $\mu_{min} = \min_{k'} \mathbb{E}[X_1^{(k')} \wedge b_1]$. Then, for any admissible policy π, we have the following regret lower bound:

$$\overline{Reg}_\pi(\tau) \geq \mu_{min} \sum_{k : d^{(k)} > 0} d^{(k)} \cdot \mathbb{E}[T_{l^*(k)}^{(k)}(\underline{n}_{\delta,\tau})] - O(1), \tag{61}$$

where $\underline{n}_{\delta,\tau} = \frac{\tau}{\mu_{max} + \delta}$ for any $\delta > 0$.

The intuition behind Lemma C.1 is as follows: each suboptimal arm decision under π yields a regret rate of $r^* - r^{(k)}(b_k^*)$. Since μ_{min} is a lower bound for the mean completion time of each epoch, the RHS of (61) yields a lower bound for the regret.

Proof. The proof follows from the identical steps as Lemma B.1 and Lemma B.2, so we will provide a proof sketch here.

In the first step, by bounding $\mathbb{E}[Rew_\pi(\tau)]$ similar to Lemma B.1, we can show the following result for all $\underline{n}$:

$$\overline{Reg}(\tau) \geq \mu_{min} \cdot \Big((\sum_k d^{(k)})\, \mathbb{E}[N_\pi(\tau) \wedge \underline{n}] - \sum_k \mathbb{E}[\underline{n} - T_{l^*(k)}^{(k)}(\underline{n})]\, d^{(k)} \Big) - 2\Phi_\pi(\tau) - O(1), \tag{62}$$

where $\Phi_\pi(\tau)$ is the overshoot function defined in (31).

In the second step, we show that the following holds under any policy that satisfies the universal optimality condition (37):

$$\mathbb{E}[N_\pi(\tau) \wedge \underline{n}_{\delta,\tau}] = \underline{n}_{\delta,\tau} - O\Big(\frac{1}{\delta^{1+\gamma}}\Big). \tag{63}$$

This result is proved by using concentration inequalities similar to the proof of Lemma B.2. Substituting (63) into (62) yields the result. □

Lemma C.1 establishes a connection between the number of suboptimal decisions under a good policy π and its regret. In the following, we provide lower bounds for the number of suboptimal decisions based on [25].

Lemma C.2. *Under Assumption 3, for any π that satisfies*

$$\mathbb{E}[T_{l^*(k)}^{(k)}(n)] = o(n^\alpha),\ \forall \alpha > 0, n \to \infty, \tag{64}$$

for any suboptimal arm k, the following lower bound holds:

$$\liminf_{n\to\infty} \frac{\mathbb{E}[T_{l^*(k)}^{(k)}(n)]}{\log n} \geq \sum_{k:d^{(k)}>0} \frac{d^{(k)}}{\inf_{\theta'\in\Theta_k^*} D_k(\theta||\theta')}, \tag{65}$$

where

$$D_k(\theta||\theta') = \mathbb{E}[\log \frac{dP_\theta^{(k)}}{dP_{\theta'}^{(k)}}(Y_1(k,b_k^*))],$$

is the Kullback-Leibler (KL) divergence between $P_\theta^{(k)}$ and $P_{\theta'}^{(k)}$.

Proof. We present a Lai-Robbins style lower bound here, which makes use of the change of measure argument as in [25]. Consider a suboptimal interrupt time indexed by $k \neq 1$ and $\theta \in \Theta_1^*$. First, by Assumption 3, we can find $\theta' \in \Theta_k^*$ such that the following holds for sufficiently small $\epsilon > 0$:

$$D_k(\theta||\theta') \leq (1+\epsilon) \inf_{\bar{\theta}\in\Theta_k^*} D_k(\theta||\bar{\theta}). \tag{66}$$

Let

$$\hat{D}_n = \sum_{s=1}^{n} \log \frac{dP_\theta^{(k)}}{dP_{\theta'}}(Y_s(k,b_k^*)),$$

be the empirical Kullback-Leibler distance. For the $\epsilon > 0$ in (66), let deterministic sequences of real numbers a_n and c_n be defined as follows:

$$\begin{aligned} a_n &= \frac{1-\epsilon}{D_k(\theta||\theta')} \log n, \\ c_n &= (1+\epsilon/2)a_n \cdot D_k(\theta||\theta'). \end{aligned}$$

Thus, we can express $\mathbb{P}(T_{l^*(k)}^{(k)}(n) \leq a_n)$ as follows:

$$\mathbb{P}(T_{l^*(k)}^{(k)}(n) < a_n) = \mathbb{P}(T_{l^*(k)}^{(k)}(n) < a_n, \hat{D}_n \leq c_n) + \mathbb{P}(T_{l^*(k)}^{(k)}(n) < a_n, \hat{D}_n > c_n). \tag{67}$$

In the following, we show that both terms on the RHS of (67) is $o(1)$ as $n \to \infty$.

In order to show that the first term on the RHS of (67) vanishes as $n \to \infty$, we make use of the change of measure along with the universal optimality of the policy. Let $A = \{T_{l^*(k)}^{(k)}(n) = n_k, \hat{D}_{n_k} \leq$

$c_n\}$ for a given deterministic sequence (c_n). Then, it is easy to show the following by using a change of measure argument:

$$\mathbb{P}(A) \leq e^{c_n}\mathbb{P}'(A),$$

where $\mathbb{P}$ and $\mathbb{P}'$ are the joint distribution of the history under θ and θ', respectively. Similarly, let $\mathbb{E}'$ be the expectation under θ'. Then, we have the following:

$$\begin{aligned}
\mathbb{P}(T_{l^*(k)}^{(k)}(n) = n_k, \hat{D}_{n_k} \leq c_n) &\leq e^{c_n}\mathbb{P}'(T_{l^*(k)}^{(k)}(n) < a_n), \\
&\leq e^{c_n}\frac{\mathbb{E}'[n - T_{l^*(k)}^{(k)}(n)] - o(n)}{n - a_n}, \\
&= o(1),\ n \longrightarrow \infty,
\end{aligned} \tag{68}$$

where the first line follows from the above change of measure argument, the second line follows from Markov inequality and the universal optimality assumption for suboptimal interrupt times, and the last line follows from the assumption that the policy is universally optimal.

In order to prove that the second term on the RHS of (67) vanishes as $n \longrightarrow \infty$, first observe that

$$\{T_{l^*(k)}^{(k)}(n) < a_n, \hat{D}_{T_{l^*(k)}^{(k)}(n)} > c_n\} \subset \{\max_{s<a_n} \hat{D}_s > c_n\}. \tag{69}$$

Let $A' = \{T_{l^*(k)}^{(k)}(n) < a_n, \hat{D}_{T_{l^*(k)}^{(k)}(n)} > c_n\}$. Then, for

$$\lambda_n = c_n - a_n \cdot D_k(\theta||\theta') = \epsilon \cdot a_n \cdot D_k(\theta||\theta')/2,$$

the following inequalities hold:

$$\begin{aligned}
\mathbb{P}(A') &\leq \mathbb{P}\Big(\max_{s<a_n} \hat{D}_s > c_n\Big), \\
&\leq \mathbb{P}\Big(\max_{s<a_n} |\hat{D}_s - s \cdot D_k(\theta||\theta')| > \lambda_n\Big) \\
&\leq \frac{a_n}{\lambda_n^2} Var\Big(\log \frac{dP_\theta^{(k)}}{dP_{\theta'}^{(k)}}(Y_s(k, b_k^*))\Big),
\end{aligned} \tag{70}$$

where the first line follows from the relation (69), and the last line follows from Kolmogorov's maximal inequality for martingales [16]. Note that $Var\Big(\log \frac{dP_\theta^{(k)}}{dP_{\theta'}^{(k)}}(Y_s(k, b_k^*))\Big) \leq 4$ for any distribution, and $\frac{a_n}{\lambda_n^2} \longrightarrow 0$ as $n \longrightarrow \infty$. Thus, the second term on the RHS of (67) vanishes as $n \longrightarrow \infty$ as well, which implies that,

$$\mathbb{P}(T_{l^*(k)}^{(k)}(n) < a_n) = 1 - \mathbb{P}(T_{l^*(k)}^{(k)}(n) > a_n) = o(1),\ n \longrightarrow \infty.$$

Hence, by Markov inequality, the following inequality holds:

$$\frac{\mathbb{E}[T_{l^*(k)}^{(k)}(n)]}{a_n} \geq 1 + o(1),\ n \longrightarrow \infty. \tag{71}$$

By using (66), we deduce that the following holds:

$$\liminf_{n\to\infty} \frac{\mathbb{E}[T_{l^*(k)}^{(k)}(n)]}{\log n} \geq \frac{1-\epsilon}{1+\epsilon} \frac{1}{\inf\limits_{\bar{\theta}\in\Theta_k^*} D_k(\theta||\bar{\theta})}, \tag{72}$$

which completes the proof. □

Substituting the lower bound for the number of suboptimal decisions in Lemma C.2 to the regret lower bound in Lemma C.2 yields the result.

REFERENCES

[1] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. 1987. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards. *IEEE Trans. Automat. Control* 32, 11 (1987), 977–982.

[2] Søren Asmussen. 2008. *Applied probability and queues*. Vol. 51. Springer Science & Business Media.

[3] 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.

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

[5] Theophilus Benson, Aditya Akella, and David A Maltz. 2010. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 267–280.

[6] Donald A Berry and Bert Fristedt. 1985. Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability). *London: Chapman and Hall* 5 (1985), 71–87.

[7] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.

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

[9] Jun Cai and José Garrido. 1999. A unified approach to the study of tail probabilities of compound distributions. *Journal of Applied Probability* 36, 4 (1999), 1058–1073.

[10] Olivier Catoni et al. 2012. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Vol. 48. Institut Henri Poincaré, 1148–1185.

[11] Nicolo Cesa-Bianchi and Gábor Lugosi. 2012. Combinatorial bandits. *J. Comput. System Sci.* 78, 5 (2012), 1404–1422.

[12] Richard Combes, Alexandre Proutiere, Donggyu Yun, Jungseul Ok, and Yung Yi. 2014. Optimal rate sampling in 802.11 systems. In *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2760–2767.

[13] Varsha Dani, Thomas P Hayes, and Sham M Kakade. 2008. Stochastic linear optimization under bandit feedback. (2008).

[14] 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.

[15] Allen B Downey. 2001. Evidence for long-tailed distributions in the internet. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*. ACM, 229–241.

[16] Robert G Gallager. 2013. *Stochastic processes: theory for applications*. Cambridge University Press.

[17] Harsh Gupta, Atilla Eryilmaz, and R Srikant. 2018. Low-Complexity, Low-Regret Link Rate Selection in Rapidly-Varying Wireless Channels. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 540–548.

[18] Allan Gut. 2009. *Stopped random walks*. Springer.

[19] András György, Levente Kocsis, Ivett Szabó, and Csaba Szepesvári. 2007. Continuous Time Associative Bandit Problems.. In *IJCAI*. 830–835.

[20] Mor Harchol-Balter. 1999. The Effect of Heavy-Tailed Job Size Distributions on Computer System Design.. In *Proc. of ASA-IMS Conf. on Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics*.

[21] Predrag R Jelenković and Jian Tan. 2013. Characterizing heavy-tailed distributions induced by retransmissions. *Advances in Applied Probability* 45, 1 (2013), 106–138.

[22] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. 2013. Online learning under delayed feedback. In *International Conference on Machine Learning*. 1453–1461.

[23] Haya Kaspi and Avishai Mandelbaum. 1998. Multi-armed bandits in discrete and continuous time. *Annals of Applied Probability* (1998), 1270–1290.

[24] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. 2010. Regret bounds for sleeping experts and bandits. *Machine learning* 80, 2-3 (2010), 245–272.

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

[26] Keqin Liu and Qing Zhao. 2011. Multi-armed bandit problems with heavy-tailed reward distributions. In *Communication, control, and computing (allerton), 2011 49th annual allerton conference on*. IEEE, 485–492.

[27] Avi Mandelbaum. 1987. Continuous multi-armed bandits and multiparameter processes. *The Annals of Probability* (1987), 1527–1556.

[28] Stanislav Minsker et al. 2015. Geometric median and robust estimation in Banach spaces. *Bernoulli* 21, 4 (2015), 2308–2335.

[29] Rajeev Motwani, Steven Phillips, and Eric Torng. 1994. Nonclairvoyant scheduling. *Theoretical computer science* 130, 1 (1994), 17–47.

[30] 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.

[31] Sidney I Resnick et al. 1997. Heavy tail modeling and teletraffic data: special invited paper. *The Annals of Statistics* 25, 5 (1997), 1805–1869.

[32] Robert Sheahan, Lester Lipsky, Pierre M Fiorini, and Søren Asmussen. 2006. On the completion time distribution for tasks that must restart from the beginning if a failure occurs. *ACM SIGMETRICS Performance Evaluation Review* 34, 3 (2006), 24–26.

[33] Peter Whittle. 1980. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)* (1980), 143–149.

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

[35] Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. 2016. Budgeted Multi-Armed Bandits with Multiple Plays.. In *IJCAI*. 2210–2216.

Received February 2019; revised March 2019; accepted April 2019