Stochastic Multi-Armed Bandits with Strongly Reward-Dependent Delays

Yifu Tang
Tsinghua University

Yingfei Wang*
University of Washington

Zeyu Zheng
University of California, Berkeley

Abstract

There has been increasing interest in applying multi-armed bandits to adaptive designs in clinical trials. However, most literature assumes that a previous patient's survival response of a treatment is known before the next patient is treated, which is unrealistic. The inability to account for response delays is cited frequently as one of the problems in using adaptive designs in clinical trials. More critically, the "delays" in observing the survival response are the same as the rewards rather than being external stochastic noise. We formalize this problem as a novel stochastic multi-armed bandit (MAB) problem with *reward-dependent* delays, where the delay at each round depends on the reward generated on the same round. For general reward/delay distributions with finite expectation, our proposed censored-UCB algorithm achieves near-optimal regret in terms of both problem-dependent and problem-independent bounds. With bounded or sub-Gaussian reward distributions, the upper bounds are optimal with a matching lower bound. Our theoretical results and the algorithms' effectiveness are validated by empirical experiments.

1 INTRODUCTION

Clinical trials are conducted to evaluate the efficacy of new treatments. In contrast with the traditional randomised clinical trials, which allocated patients to the two treatments uniformly at random throughout the trial, there has been growing interest in adopting response-adaptive randomisation (RAR), in which the randomisation probabilities change during the trial as patient responses are observed. This dynamic design, aimed at enhancing efficiency and ensuring that more patients are assigned to the better performing treatments, has become an attractive method for patient allocation (Pallmann et al., 2018; Zhang and Rosenberger, 2007; Williamson et al., 2022).

While ensuring a clinical trial has enough power to identify meaningful differences at its end is important, we must also prioritize the well-beings of the patients throughout the trial. Rising to this opportunity, multi-armed bandits (MAB) have emerged to be the idealized mathematical decision framework for response-adaptive clinical trials. As such, the last decade has witnessed tremendous research efforts in designing efficient (stochastic) bandit algorithms to correctly identify the best treatment while treating patients as effectively as possible during the trial (Aziz et al., 2021; Varatharajah and Berry, 2022; Atan et al., 2019; Zhang and Rosenberger, 2007; Arya and Yang, 2020; Zhou et al., 2019a), although the scope of multi-armed bandits is much more general.

However, in most multi-armed bandit settings it is assumed that the reward of treatment allocation is immediately available before the next patient arrives. This is not realistic since in most cases the treatment effect is seen at some delayed time after the treatment is provided. The challenges introduced by delayed outcomes in oncology and cancer treatments where the therapeutic effect takes time to manifest are well-recognized in the literature (Arya and Yang, 2020; Eick, 1988a,b; Williamson et al., 2022). The inability of most response-adaptive designs to account for delay has long been cited as one of the major hindrances to their practical application (Shrestha and Jain, 2021; Williamson et al., 2022).

The importance of considering delays is highlighted by literature in recent years, while most considerations were motivated by reward delays in advertisement and news article recommendations (Vernade et al., 2017; Joulani et al., 2013; Vernade et al., 2020; Zhou et al.,

*Correspondence to: Yingfei Wang <yingfei@uw.edu>.
Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

2019b; Gael et al., 2020). As such, the delay was assumed either to be fixed, or stochastic but *reward-independent*, i.e. sampled from an unknown delay distribution that can depend on the chosen arm, but *not* on the stochastic reward received on the same round. In many clinical settings, however, it is seldom the case that the individual medical outcomes will be available later at random points in time, regardless of the effectiveness of the treatment.

The more challenging setting of *reward-dependent* delays was not explicitly addressed previously in the bandit literature, expect for the work by Lancewicki et al. (2021). More critically, most of the researchers and physicians for oncology and cancer treatment use survival analysis to measure the outcome of a treatment. Among all the survival measures, progression-free survival (PFS) is defined as the number of days after the treatment until a disease progression or death, and is widely used as an event-driven surrogate measure of clinical benefits (Driscoll and Rixe, 2009; Hari et al., 2018; Marshall et al., 2016). In this case, we can observe the survival outcome only until disease worsens, in the sense that the "delays" in observing the PFS are the same as the PFS itself, rather than external stochastic noise. In contrast to reward-independent case, the *observed* survival outcomes can give a biased estimation of the true rewards as shorter PFS will be observed earlier. Meanwhile, a unique feature of survival analysis is that typically not all patients experience the event (e.g., death or disease progression) by the end of the observation period, so the actual survival responses for some patients are unknown. This phenomenon, referred to as censoring, must be accounted for in the analysis. Hence, when a substantial portion of survival responses remain unobserved, and when reward and delay can be arbitrarily large, the expected *observed* reward of the best arm might be much smaller than that from a sub-optimal arm, which creates significant challenges in learning.

1.1 Our Contributions

Motivated by the existing gap between the theory of response-adaptive randomisation (which is abundant with clinical trial design proposals in the setting of immediate responses) and clinical practice (in which survival responses are typically delayed), we propose a novel stochastic bandit formulation that maximizes the cumulative expected rewards r_t of adaptively chosen arms a_t, given stochastic delays $d_t = \max(0, \lceil r_t \rceil)$ in observing the reward r_t, which creates a (nearly) perfect correlation between the stochastic reward r_a and the delay distribution τ_a[1].

[1] $\lceil x \rceil$ denotes the minimum integer that is $\geq x$.

As progression-free survival (PFS) is naturally positive, we first consider general reward distributions with no assumptions other than $r_a > 0$ and $\mathbb{E}[r_a] < \infty$, for any arm a. We design an anytime Censored-UCB algorithm which achieves $\mathcal{O}\left(\sum_{i \neq i^*} \frac{\log(T) \log\log(T)^2}{\Delta_i}\right)$ expected regret and a problem-independent (i.e. gap-free) regret bound of $\mathcal{O}\left(\sqrt{KT \log(T)} \log\log(T)\right)$, that match the bounds of the standard non-delayed setting up to a log-logarithm factor. The additional log-logarithm factor can be attributed to potentially heavy-tailed reward/delay distributions and the fact that long delays are associated with larger rewards due to their strong dependency.

We then consider important special cases with bounded reward/delay distributions, and sub-Gaussian reward/delay distributions. We provide a refined analysis that achieves problem-dependent regret bounds of $\mathcal{O}(\log(T))$, with an additive increase of $\sum_{i \neq i^*} \Delta_i$ in the regret for the standard non-delayed stochastic MAB. We show that this bound is optimal under bounded rewards, by presenting a matching lower bound. We also present the first problem-independent bound of $\mathcal{O}\left(\sqrt{KT \log(T)}\right)$ under reward-dependent delays that matches the bounds of the standard non-delayed MAB.

We would like to note here that our settings and herein the results are new compared to the literature. To the best of our knowledge, the only studies (Lancewicki et al., 2021; Gael et al., 2020) that propose provable algorithms for *reward-dependent* setting rely on the assumption that the rewards are bounded in [0,1], with additional assumptions in Gael et al. (2020) such as the delays are bounded by an α-pareto distribution.

1.2 Related Literature

Recent research has explored the challenges of learning in bandit feedback with delays across various contexts in the literature. Most of the work assumes *reward-independent* delays, i.e. the reward r_t and delay d_t are independent from each other. Among the related studies, Dudík et al. (2011) is the first to consider delays in stochastic MAB with fixed delay d. Mandel et al. (2015) follows the work of Joulani et al. (2013) and allows for some unknown, arbitrary process that generates delay times. Similar to Joulani et al. (2013), the delay is assumed to be bounded, and the upper bound is consistent with the classical MAB problems.

There are other works studying i.i.d. stochastic delays. For arm-independent stochastic delays, contextual bandits are considered in Zhou et al. (2019b). Vernade et al. (2020) proposes learning algorithms for linear Bandits in which delays are only partially

observable. Zhou et al. (2019b) investigates generalized linear contextual bandits in the presence of i.i.d. stochastic delays. Pike-Burke et al. (2018) proposes a variant of delayed bandits with aggregated anonymous feedback, under the assumption that the expected delay is bounded and known to the learner. Arm-dependent stochastic delays have been investigated by Gael et al. (2020) in the setting of stochastic bandits, Lancewicki et al. (2021) with less assumptions with unrestricted delay distributions, and Arya and Yang (2020) under a multi-armed bandit problem with covariates. A more recent work by Shi et al. (2023) devotes to establishing valid statistical inference to quantify the uncertainty of learned policies under multi-armed bandits with arm-dependent delayed feedback. Keyvanshokooh et al. (2019) proposes a contextual multi-armed bandit model that discusses feedback under (reward-independent) constant delays and stochastic delays and highlights the challenges posed by delayed feedback in medical contexts.

Much less attention has been paid on *reward-dependent* stochastic delays. Vernade et al. (2017) considers delayed Bernoulli bandits where it is impossible to decide whether the conversion is 0 or if conversion is 1 but the observation is delayed. However, this work requires complete knowledge of the delay distribution. Lancewicki et al. (2021) more explicitly proposes the unrestricted delay distributions, in which the stochastic delay at each round and the reward are drawn from a joint distribution. While there is no assumption on the delay distributions, the reward distributions are assumed to be bounded in [0,1]. Although with a slightly different focus, Zimmert and Seldin (2020); Gyorgy and Joulani (2021); Van Der Hoeven and Cesa-Bianchi (2022) consider non-stochastic bandits under arbitrary or arm-dependent delays and propose algorithms achieving regret bounds that depend on the total delay D in T rounds.

On the other hand, in the literature on clinical trials, most RAR procedures hinge on the limiting assumption that when a treatment must be assigned for a newly enrolled patient, the outcomes of previously treated patients must be fully assessed. However, as survival analysis and survival measures continue to be well acknowledged as the primary assessment to approve new drugs, design or interpret clinical trials, intrinsic delay of survival outcomes can prohibit their implementation in practice. Ryeznik et al. (2012) describes the inability to account for delay as "a major stumbling block in implementing adaptive designs", and Rosenberger et al. (2012) lists it as one of the main criticisms of response-adaptive randomisation.

Several research attempts have focused on simplified problems. For example, Eick (1988a,b); Wang (2002) examines delayed feedback within a two-armed clinical trial, where the distribution of one arm is assumed to be known. Most methodological papers discussing response-adaptive randomisation procedures assume a fixed delay (e.g. Langenberg and Srinivasan (1982); Chick et al. (2017, 2022)), with recent work (Williamson et al., 2022) identifying the existing gap for Bayesian response-adaptive randomisation procedures by considering a two-armed Bernoulli bandit with either fixed and/or arm-independent stochastic delays. Xu and Yin (2014) and Zhang and Rosenberger (2007) propose optimal allocation schemes derived for arm-independent exponential and/or Weibull distributed response delays.

While empirical and medical evidence have suggested the use of bandit formulation for response-adaptive randomisation with PFS as response (for example, Zhou et al. (2019a) have shown the effectiveness of a Bayesian contextual bandit algorithm for treating an incurable cancer on 803 patients recruited in 2000-2017 with 3762 total revisits and average PFS of 163 days (median = 93)), nevertheless, there is no response-adaptive randomisation literature considered *reward-dependent* delays for general reward/delay distributions, and our work aims to fill the gap.

2 PROBLEM SETTING

Notations. We denote $\min\{a, b\}$ as $a \wedge b$. For a positive integer n, denote $[n]$ to be the set $\{1, 2, \ldots, n\}$. For any real number x, $\lceil x \rceil$ denotes the minimum integer that is not less than x, while $\lfloor x \rfloor$ denotes the maximum integer that does not exceed x. For two real-valued functions f and g, if there exists constants c, N, such that $f(t) \leq cg(t)$ for all $t \geq N$, we write this as $f = \mathcal{O}(g)$, and $g = \Omega(f)$. Finally, we denote $\mathbb{I}(\mathcal{E})$ as the indicator of the event $\mathcal{E}$.

Suppose there are $K > 1$ arms or treatments in the action set $\mathcal{A}$. Each arm $i \in \mathcal{A}$ is associated with a reward distribution ν_i and a delay distribution τ_i. The reward distribution can have unbounded support and the delay distribution is supported on $\mathbb{N} = \{0, 1, ...\}$. We denote by μ_i the (unknown) expected reward of arm i, $\mu^* = \mu_{i^*} = \max_j \mu_j$, and $\Delta_i = \mu^* - \mu_i$.

At each round t (e.g. for current patient t), the learner chooses an arm (e.g. treatment) $a_t \in \mathcal{A}$ and get reward $r_t(a_t)$ sampled i.i.d. from ν_{a_t}. Unlike the standard MAB setting, the learner does not immediately observe $r_t(a_t)$ at the end of round t. Instead, the tuple $(a_t, r_t(a_t))$ is received at the end of round $t + d_t(a_t)$.

The reward-dependent delay model. Motivated by the clinical trials setting where the reward $r_t(a_t)$ can measure the survival time after a treatment a_t is

given to the current patient t, in this work, we consider a special form of the stochastic delays $d_t(a_t)$ that are dependent on the reward $r_t(a_t)$ in the same round. For example, one commonly used survival measure is progression-free survival (PFS), which is defined as the number of days after the treatment, that a patient lives with the disease (such as cancer) but it does not get worse. The identification of "progression" generally involves imaging techniques (e.g. plain radiograms, CT scans, MRI) or the toxicological characteristics of the treatments in the trial. With this being said, if the progression-free survival is $r_t(a_t)$, which are positive integers corresponding to the number of days, the tuple $(a_t,\ r_t(a_t))$ can only be observed after $d_t(a_t) = r_t(a_t)$ rounds until the identification of disease progression.

Proposed more generally, for general reward distributions ν_i, if the sampled reward $r_t < 0$, the reward will be revealed immediately without delay. If the sampled reward $r_t \geq 0$, the delay d_t in receiving the response is the minimum integer that is no less than r_t. In other words, $d_t = \max(0, \lceil r_t \rceil)$ in observing the reward r_t, which creates a (nearly) perfect correlation between the stochastic reward distribution ν_i and the delay distribution τ_i.

Performance criterion. In most bandit problems, the regret is the cumulative loss due to not playing an optimal action. The performance of the learner in our setting is thus measured as usual by the expected pseudo-regret that considers the loss of all *generated* rewards, regardless of whether the reward is received before horizon T. Formally,

$$\begin{aligned}\mathbb{E}[R_T] &:= \max_i \mathbb{E}\left[\sum_{t=1}^{T} \nu_i\right] - \mathbb{E}\left[\sum_{t=1}^{T} \nu(a_t)\right] \\ &= T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \nu(a_t)\right].\end{aligned}$$

3 OUR ALGORITHM: CENSORED UPPER CONFIDENCE BOUND

At each round t, if $s+d_s < t$ (reward occurred in round s is available at the beginning of round t), (s, a_s, r_s) are completely *observed* by the learner. In the presence of delayed feedback, however, the sample mean of completely observed rewards is no longer an unbiased estimator for μ_i, as the expectation of observed reward can be different from the actual expected reward, e.g. smaller reward will be observed earlier.

The essential idea of our proposed algorithm is to use censored information of the reward within m steps after the arm was played, to augment the set of completely observed tuples. Namely, for a given positive integer m, for any $t > s + m$, if delay $d_s \leq m$, the tuple (s, a_s, r_s) is included in the history $\mathcal{H}_t$ of the observed information. If $d_s > m$, given the dependency between the reward and the delay distribution, while r_s may not be observed due to delay, we know that $r_s > m$. Hence the tuple (s, a_s, m) is added into the history $\mathcal{H}_t$ of information. Taking together, we define

$$\begin{aligned}\mathcal{H}_t(m) &= \{(s, a_s, r_s) | \forall s, s+m < t, d_s \leq m\} \\ &\quad \cup \{(s, a_s, m) | \forall s, s+m < t, d_s > m\} \\ &= \{(s, a_s, r_s \wedge m) | \forall s, s+m < t\}.\end{aligned}$$

as the information (filtration) available at the beginning of round t for the learner to choose an arm.

Denoting the number of pulls of arm i up to time $t-m$ as $N_t(m,i) := \sum_{s=1}^{t-m-1} \mathbb{I}\{a_s = i\}$, we construct the estimator as follows:

$$\hat{r}_t(m,i) := \frac{\sum_{s=1}^{t-m-1} (r_s \wedge m)\mathbb{I}\{a_s = i\}}{N_t(m,i)}. \tag{1}$$

The idea is that if we keep increasing m (which might be a function of the round t, i.e., $m = \lfloor m(t) \rfloor$), we have $\lim_{m\to\infty} \mathbb{E}[\nu_i \wedge m] = \mathbb{E}[\nu_i]$, based on the Dominated Convergence Theorem (see appendices for proof). Thus the estimation of $\mathbb{E}[\nu_i \wedge m]$ will asymptotically converge to the real mean rewards.

3.1 Algorithm

We first consider general reward distributions with potentially unbounded support, under the only assumption of finite expectation, i.e. $\mathbb{E}[\nu_i] < \infty$ for any $i \in \mathcal{A}$, which includes exponential and Weibull distributions that are frequently used for survival analysis, Poisson distributions, and other heavy-tailed distributions. As survival measures are naturally positive, we start our discussion with $\nu_i > 0$.

We first define a censored upper confidence bound based on the following concentration bound, with the proof provided in the appendix.

Proposition 1. *For any non-random integers* $1 < m < t \leq T$, *and* $\delta \in (0, \frac{1}{2})$, *we have*

$$\mathbb{P}\left(|\hat{r}_t(m,i) - \mathbb{E}[\nu_i \wedge m]| \leq m\sqrt{\frac{2}{N_t(m,i)} \log\left(\frac{2T}{\delta}\right)}\right) \geq 1-\delta.$$

By applying the union bound for $0 \leq t \leq T$ and $i \in \mathcal{A}$, we define the $m(t)$-censored-UCB index as

$$UCB_i^{m(t)}(t,T) = \hat{r}_t\Big(m(t), i\Big) + m(t)\sqrt{\frac{2\log\left(2K^2T^3\right)}{N_t\Big(m(t), i\Big)}},$$

where $m(t) = o(t)$ is a censoring function that may take the forms of, for example, $\log(t)$ and $\log\log(t)$.

Algorithm 1 presents the censored-UCB algorithm for known horizon T. To initialize Algorithm 1, one should begin the UCB index when $t - m(t) - 1 > K$ so that each $\nu_i \wedge M$ should have at least one sample to build the estimator $\hat{r}_t\Big(m(t), i\Big)$. We denote l as the number of initialization rounds that are required. One should note that l is not random, and only depends on the censored function $m(\cdot)$ and the number of arms. If $m(t)$ is set to be $\log(t)$ or $\log\log(t)$, we can set $l = 2$.

To get an anytime algorithm, we utilize the doubling trick and select several nodes $l < T_1 < T_2 < \cdots$ as the updating points. We follow the standard exponential doubling trick (Besson and Kaufmann, 2018) such that partition $T_1 < T_2 < \cdots$ is defined to be

$$T_s = \left\lfloor \frac{T_0}{a} a^{b^s} \right\rfloor, s \in \mathbb{N}_+ \tag{2}$$

for some $a, b > 1, T_0 \in \mathbb{N}_+$. Suppose that Alg_T represents a sub-routine (e.g. Algorithm 1) with a known horizon T, algorithm 2 presents the anytime censored-UCB algorithm with unknown horizon T. We note here that the adoption of doubling trick is from a theoretical standpoint and a refined regret bound can be obtained by an upper confidence bound that only depends on the current time t.

Algorithm 1: $m(t)$-Censored-UCB algorithm

Input: Horizon T, censoring function $m(t)$, initialization rounds l

Pull each arm l times

for $t \geq lK + 1$ **do**

  Compute $N_t\big(m(t), i\big)$ and $\hat{r}_t\big(m(t), i\big)$ for all i

  Compute upper confidence bound $UCB_i^{m(t)}(t, T)$ for all $i \in \mathcal{A}$

  Select $a_t \leftarrow \arg\max_i UCB_i^{m(t)}(t, T)$

end

Algorithm 2: Anytime algorithm with unknown horizon T

Input: The updating nodes $T_1 < T_2 < \cdots$, the sub-algorithm Alg_T with any given T

$i \leftarrow 0$, initialize the algorithm $\text{Alg}^{(0)} = \text{Alg}_{T_0}$;

for $t = 1, 2, 3, \cdots$ **do**

  if $t > T_i$ **then**

    $i \leftarrow i + 1$

    Initialize algorithm $\text{Alg}^{(i)} = \text{Alg}_{T_i - T_{i-1}}$

  end

  Play algorithm $\text{Alg}^{(i)}$: the arm a_t is selected the same as that in round $t - T_i$ under $\text{Alg}^{(i)}$

end

3.2 Regret Analysis

The regret analysis presented in this section is derived under Algorithm 2. The results for known horizon T are discussed in supplementary materials.

For any real number $c > 1$ and sub-optimal arm i, we define $d_{i,i^*}(c)$ as

$$0 \vee \inf\left\{ x \in \mathbb{R} : \forall m \geq x, \mathbb{E}[\nu_{i^*} \wedge m] - \mathbb{E}[\nu_i \wedge m] \geq \frac{\Delta_i}{c} \right\}$$

which means the smallest value in $\mathbb{R}_+ \cup \{0\}$, truncating on which we can lead to an at least $\frac{\Delta_i}{c}$ gap between arm i and the optimal arm. Such value must exist since for any arm i, $\mathbb{E}[\nu_i \wedge m] \to \mathbb{E}[\nu_i]$. Thus, for some given c, $d_{i,i^*}(c)$ is a problem-dependent parameter.

Similarly, we define

$$d_{i,i^*} = \min\left\{ j \in \mathbb{N} : \mathbb{E}\left[\nu_i \wedge k\right] < \mathbb{E}\left[\nu_{i^*} \wedge k\right], \forall k \geq j \right\},$$

which can be viewed as $d_{i,i^*}(\infty)$.

We note here that, in general, d_{i,i^*} cannot be merely determined by the mean and the variance of ν_i and ν_{i^*}. For example, we let $\nu_{i^*} \sim \mathcal{N}(1.1, 1)$, $\nu_1 \sim Poisson(1)$, $\nu_2 \sim \mathcal{N}(1, 1)$, where ν_1 and ν_2 have the same mean and variance. However, by definition, $d_{2,i^*} = 0$, $d_{1,i^*} > 0$. If the reward distributions are limited to a double-parameter exponential family, whose probability density functions are defined in

$$\left\{ f(x) = e^{(x-\mu)^2/2\sigma^2} / \sqrt{2\pi\sigma^2} : \mu, \sigma \in \Theta \right\},$$

then since ν_i's are characterized by mean and variance,

$$d_{i,i^*} = \inf\left\{ k > 0 : \int_m^\infty (F_{i^*}(t) - F_i(t))dt \leq \Delta_i, \forall m \geq k \right\}$$

is determined by $\Delta_i, \mu_i, \sigma_i$.

Theorem 1. *Suppose there exists a constant $c_0 > 0$ such that for all $i \in \mathbb{N}_+$, $\frac{m(T_i)}{m(T_{i-1})} \leq c_0$, then for any $c > 1$ and $T > \max\{C_0, K\}$, the expected regret of Algorithm 2 satisfies*

$$\begin{aligned}\mathbb{E}[R_T] &\leq \sum_{i \neq i^*} \frac{C_1 c^2 m(T)^2}{\Delta_i} \log(T) + \sum_{i \neq i^*} C_2 m(T) \log\log(T) \Delta_i \\ &\quad + C_3 \log(\log(T)) \sum_{i \neq i^*} m^{-1}(d_{i,i^*}(c)) \Delta_i + C_4 \\ &= \mathcal{O}\left(\sum_{i \neq i^*} \frac{m(T)^2 \log(T)}{\Delta_i} \right),\end{aligned}$$

where $C_i = C_i(a, b, c_0, T_0), i = 0, 1, 2, 3, 4$ are some universal constants.

The condition $\frac{m(T_i)}{m(T_{i-1})} \leq c_0$, is to say that $m(\cdot)$ is not increasing very fast. For logarithm and iterated logarithm censoring functions, this condition can be satisfied by $c_0 = b$, $T_0 > a, b > 1$. Meanwhile, $\log\log(T)$ in

the second and the third term comes from the process of exponential doubling trick and the regret bound for know horizon T does not include this term.

Proof of Theorem 1 (sketch). We first derive the upper bound for a given budget T, which is of order $\sum_{i\neq i^*} M^2(T)\log(T)/\Delta_i$. Consider the last time the agent pulls an arm i that is sub-optimal. The Algorithm ensures that $UCB_{i^*}^{m(t)}(t,T) \leq UCB_i^{m(t)}(t,T)$ at this time, leading to the following bound for $\mathbb{E}(\nu_{i^*} \wedge m(t)) - \mathbb{E}(\nu_i \wedge m(t))$, i.e.,

$$\mathbb{E}(\nu_{i^*} \wedge m(t)) - \mathbb{E}(\nu_i \wedge m(t)) \leq 2m(t)\sqrt{\frac{\log(2K^2T^3)}{N_t(m(t),i)}}.$$

If at this time, t is greater than $m^{-1}(d_{i,i^*}(c))$, then the left hand side will be greater than $\frac{\Delta_i}{c}$ thus leading to an upper bound for $N_t(m(t),i)$:

$$N_t(m(t),i) \leq C'\frac{c^2m^2(t)\log(T)}{\Delta_i^2},$$

for some universal constants C', using the fact that $K < T$. Otherwise, $t < m^{-1}(d_{i,i^*}(c))$, then the total number of pulls on the arm i will not exceed t, and thus the reward contributed by this arm is bounded by $m^{-1}(d_{i,i^*}(c))\Delta_i$. The known horizon upper bound is obtained by taking the pulls from $t-m-1$ to t into consideration, namely, a regret of order $m(T)\sum_{i\neq i^*}\Delta_i$. Then, according to the doubling trick, the regret at time T is accumulated by

$$R_T = \sum_{j=1}^{L_T} R_{T_j - T_{j-1}}(\mathrm{Alg}_{T_j - T_{j-1}})$$

where $L_T = \min\{j : T_j > T\}$. Since the summation $\sum\log(T_i - T_{i-1})$ conserves the order of $\log(T)$, the doubling trick conserves the order. For the low-order terms, $L_T = O(\log\log T)$ implies the contribution of which cannot exceed the leading terms. □

Remark 1. Although the term $m(\cdot) = o(t)$ can be arbitrarily small, this is not a free lunch, since the constant involves a term $m^{-1}(d_{i,i^*}(c))$. It will be infinitely large if one chooses $m(t)$ to be constant. As a result, this includes a trade-off. If we choose the censoring function to be $m(t) = \log\log(t)$, the regret is bounded by $\mathcal{O}\left(\sum_{i\neq i^*}\frac{\log\log(T)^2\log(T)}{\Delta_i}\right)$ that matches the $\mathcal{O}\left(\sum_{i\neq i^*}\frac{\log(T)}{\Delta_i}\right)$ regret bound of the classic MAB without delays up to a log-logarithm factor. The slight gap can be attributed to potentially heavy-tailed distributions, where $m^2(t)$ gives an upper bound of the arbitrarily large second moment of the censored reward $\nu_i \wedge m(t)$ up to time t. The term $m^{-1}(d_{i,i^*}(c))\Delta_i$ corresponds to the number of pulls we use to distinguish the difference between arms. Such a term can be unexpectedly large if the distributions are irregular when the sub-optimal arm i dominated the truncated expected rewards $\mathbb{E}(\nu_i \wedge N)$ until a very large N, which is intuitively called the "stick-together" arms. Another term, $m(t)\sum_{i\neq i^*}\Delta_i$ corresponds to the regret because of the information that has been censored, with $m(t)$ as the cost we pay to process that censored information.

The above regret is problem-dependent in the sense that it is fully specified by the reward/delay distributions of the arms. When $\min_i \Delta_i$ is small and/or $d_{i,i^*}(c)$ is large, the problem-dependent constants can be usually large. Hence, we also notice that it is valuable to derive problem-independent bounds (e.g. distribution-free) as follows which are bounds on the worst-case expected regret as a function of the number of arms K and time T (Degenne and Perchet, 2016; Joulani et al., 2013; Pike-Burke et al., 2018).

Theorem 2 (Problem-independent)**.** *If there exists some constant c_0 such that $\frac{m(T_{i+1})}{m(T_i)} \leq c_0$, and if we further assume additional uniform integrability on ν_i, i.e. $\mathbb{E}\left[m^{-1}(x)(\nu_i - \nu_i \wedge x)\right] < G_1$ and $\mathbb{E}(\nu_i) \leq G_2 < \infty$ for some constants G_1 and G_2, then Algorithm 2 incurs an expected regret at most*

$$\begin{aligned}\mathbb{E}[R_T] &\leq C_1 m(T)\sqrt{KT\log(T)} + C_2 \\ &= \mathcal{O}\left(m(T)\sqrt{KT\log(T)}\right),\end{aligned}$$

for $T > C_3$, where $C_i = C_i(a,b,T_0,G_1,G_2), i = 1,2,3$ are universal constants.

The uniform integrability assumption generally means that the increasing speed of $m^{-1}(\cdot)$ is slower than the decaying speed of the tail probability. For example, if $m(t) = \log t$ while ν_i is assumed to have sub-Gaussian distributions, the condition is satisfied.

4 TIGHTER REGRET BOUNDS FOR SPECIAL CASES

In this section, we show that under some common and mild assumptions on the reward distribution, we can obtain tighter regret guarantees. We consider two settings: reward distribution with bounded support, and sub-Gaussian reward distributions.

4.1 Rewards with Bounded Support

Without loss of generality, we assume that the reward distribution ν_i is supported in $[0, M]$ and known to the learner, with the generalization to any bounded case $[M_1, M_2]$ provided in the appendix.

In this case, for time $t > M$, the rewards from the time before $t - M - 1$ must be observed by time t. Hence

we propose to use the following estimator

$$\hat{r}_t(i) = \frac{\sum_{s=1}^{t-M-1} r_s \mathbb{I}(a_s = i)}{N_t(i)},$$

where $N_t(i) = \sum_{s=1}^{t-M-1} \mathbb{I}(a_s = i)$. Note that M remains constant in this setting. Since all the outcoming rewards are observed after at most M time points, this estimator is unbiased. Similarly, the upper confidence bound is constructed as

$$UCB_i(t,T) = \hat{r}_t(i) + M\sqrt{\frac{2\log\left(2K^2T^3\right)}{N_t(i)}}.$$

For completeness, we present the pseudo-codes of bounded rewards in Algorithm 3. Algorithm 2 can be used to obtained an anytime algorithm by adopting Algorithm 3 as the sub-routine Alg_T.

Algorithm 3: Censored-UCB algorithm for known horizon T under bounded reward setting

Input: Horizon T, M
Pull each arm once
Random pull the arms until $t > M$
for $t \geq \max\{K, M+1\}$ **do**
  Compute the estimation $\hat{r}_t(i)$ for $i \in \mathcal{A}$
  Compute $UCB_i(t,T)$ for $i \in \mathcal{A}$
  Select $a_t \leftarrow \arg\max_i UCB_i(t,T)$
end

We get the following instance specific and problem-independent regret bounds.

Theorem 3. *The expected regret of anytime Algorithm 3 can be upper bounded by*

$$\begin{aligned}\mathbb{E}[R_T] &\leq C_1 \sum_{i \neq i^*} \frac{M^2 \log(T)}{\Delta_i} \\ &\quad + C_2 M \log(\log T) \sum_{i \neq i^*} \Delta_i + C_3 \\ &= \mathcal{O}\left(\sum_{i \neq i^*} \frac{\log(T)}{\Delta_i}\right),\end{aligned}$$

$T > C_4$, with universal constants $C_i = C_i(a, b, T_0), i = 1, 2, 3, 4$.

We remark that $\log\log(T)$ comes from the process of exponential doubling trick and the regret bound for know horizon T does not include the $\log\log(T)$ term. It is worth noting that the upper bound by Lancewicki et al. (2021) depends on $d_i(q_i)$ that represents quantiles of the arms' distribution, while our bound only depends on Δ_i.

Theorem 4 (Problem-independent)**.** *For any problem instance, the expected regret of Algorithm 3 satisfies,*

$$\begin{aligned}\mathbb{E}[R_T] &\leq C_1 M\sqrt{KT\log(T)} + 2KM^2 \\ &= \mathcal{O}(\sqrt{KT\log(T)}),\end{aligned}$$

for universal constant C_1, and $\forall T > \max(K, M)$.

It worth noting here, the best-known problem-independent bound for the expected regret of classic (non-delayed) UCB1 is $\mathcal{O}(\sqrt{KT\log(T)})$, together with $\mathcal{O}(\sqrt{KT\log(T)})$ for Thompson sampling using Beta priors (Auer et al., 2002; Agrawal and Goyal, 2017; Bubeck et al., 2012). Meanwhile, Joulani et al. (2013) shows that for *reward-independent* delay distributions with a finite expected delay, the worst case scales with $\mathcal{O}(\sqrt{KT\log(T)} + K\mathbb{E}[\tau])$. Nevertheless, our results show that under our *reward-dependent* delay with bounded support, the upper bound matches with the order of the classical MAB problems, and recovers the result by Joulani et al. (2013) up to an M factor which reflects the upper bound of the reward distribution, saying that the price to pay for the delay in receiving the observations is negligible.

4.2 Sub-Gaussian Reward Distributions

In this section, we consider the broadly adopted setting in MAB with sub-Gaussian reward distributions. Without loss of generality, we assume that all arms are 1-sub-Gaussian defined as follows,

Definition 1. *An arm with stochastic reward X is said to be 1-sub-Gaussian, if for all $t > 0$,*

$$\mathbb{P}\left(X - \mathbb{E}X \geq t\right) \leq e^{-\frac{t^2}{2}}, \mathbb{P}\left(X - \mathbb{E}X \leq -t\right) \leq e^{-\frac{t^2}{2}}$$

Consequently, $\forall \lambda \in \mathbb{R} : \mathbb{E}\left(e^{\lambda(X - \mathbb{E}X)}\right) \leq e^{\lambda^2/2}$.

We modified the UCB index in Algorithm 1 as follows,

$$UCB_i^{m(t)}(t,T) = \hat{r}_t(m(t),i) + \sqrt{\frac{4}{N_t(m(t),i)}\log\left(e^2K^2T^3\right)}.$$

The key idea is that when ν_i are assumed to have sub-Gaussian distributions, the distribution of $\nu_i \wedge m$ is "nearly sub-Gaussian", by which we can obtain a refined concentric inequality of $\hat{r}_t(m(t),i)$. Using the censored function $m(t) = \log t$, we have the following regret bounds with proofs in the appendix.

Theorem 5. *The anytime Censored-UCB algorithm with $m(t) = \log t$ leads to an expected regret, for $T > C_0(a, b, T_0)$ and any $c > 1$,*

$$\begin{aligned}\mathbb{E}[R_T] &\leq C\left(\Delta_i + \frac{1}{\Delta_i}\right)\log(T) + 2\sum_{i \neq i^*} C_i \log\log(T) \\ &= \mathcal{O}\left(\left(\Delta_i + \frac{1}{\Delta_i}\right)\log(T)\right),\end{aligned}$$

in which $C = 128c^2b^2/(b-1), C_i = e^{d_{i,i^*}(c)}\Delta_i(4 + 2/\log b)$, *and* $C_0(a, b, T_0)$ *is some constant only related to* a, b, T_0.

Remarkably in the proposed upper bound, the first part $\sum_{i\neq i^*}\frac{\log T}{\Delta_i}$ is coming from learning, while the second part $\sum_{i\neq i^*}\Delta_i \log T$ comes from the delay $m(T) = \log(T)$. Nevertheless, this matches the $\mathcal{O}(\log(T))$ regret bound of non-delayed MAB.

Notably, again, this is not a free lunch. In our setting, the discrepancy between the two arms might be so tiny that the learner cannot distinguish it in a short period of time. Thus, one will need the problem-dependent constant C_i to control such circumstances.

Theorem 6 (Problem-independent)**.** *For any problem instance satisfying 1-sub-Gaussian reward distributions and the assumption that the expected reward is bounded by some constant* $G > 0$ *uniformly, the expected regret of the anytime Censored-UCB algorithm satisfies*

$$\mathbb{E}[R_T] \leq 16\sqrt{7KT\log(T)} + C_G K \log(T),$$

where $C_G = (11G + 24)e^{2G+8}$.

5 LOWER BOUND

In this section, we investigate the problem-dependent lower bound of the expected regret. Consider that there are two arms and i^* is the optimal. Recall that

$$d_{i,i^*} = \min\left\{j \in \mathbb{N} : \mathbb{E}\left[\nu_i \wedge k\right] < \mathbb{E}\left[\nu_{i^*} \wedge k\right], \forall k \geq j\right\},$$

which is the earliest time m, truncated by any time after which, the optimal arm i^* will always have a greater mean reward than the sub-optimal arm i. The following theorem tells us that the learner has to spend an order of d_{i,i^*} steps to distinguish the differences between the arms.

Theorem 7. *Let* $K = 2$, *and* Δ *be the sub-optimal gap, if an algorithm ALG under delays guarantees a regret bound* T^α *for all instances, then there exists a problem instance with sub-optimal gap* Δ, *for this instance, it must suffer an expected regret of*

$$\mathbb{E}[R_T^{ALG}] = \Omega\left(\frac{(1-\alpha)\log(T)}{\Delta} + d_{i,i^*}\Delta\right).$$

The lower bound term includes two terms, in which the first term comes from the standard multi-armed bandit problem. For the second term, assume that the optimal arm has a distribution $\mathbb{P}(\nu_{i^*} = 0) = 1 - p$ and $\mathbb{P}(\nu_{i^*} = M + \Delta/p) = p$, while the sub-optimal arm i has a distribution $\mathbb{P}(\nu_i = 0) = 1 - p$ and $\mathbb{P}(\nu_i = M) = p$. In this case, $M = d_{i,i^*}$. Before time M, there is no means to distinguish the subtle between the two arms, thus the regret is of order $M\frac{\Delta}{2}$, i.e., $d_{i,i^*}\Delta$.

We would like to point out that under known horizon T, Algorithm 3 achieves a regret upper bound of

$$\mathbb{E}[R_T] \leq \sum_{i\neq i^*}\frac{CM^2}{\Delta_i}\log(T) + 2M\sum_{i\neq i^*}\Delta_i,$$

with a universal constant C. This bound is optimal as it matches the lower bound w.r.t the additive increase.

6 EXPERIMENTS

We conduct various numerical experiments to illustrate our theoretical results, and investigate the effect of the delay on the performance of the algorithms. For each experimental setting, the results are averaged over 50 runs, and the average cumulative regret (with 95% confidence interval) is depicted in Figure 1.

We compare our Censored-UCB algorithm to the naive extension of the UCB algorithm in which the rewards are taken into account only if it is observed (Naive UCB), non-delayed UCB which assumes no delay in observing the feedback, and OPSE (for general reward-dependent delays) (Lancewicki et al., 2021). Note that in OPSE, it assumes that all missing samples have the maximal/minimal reward in estimating the upper/lower confidence bound, and hence OPSE is applicable only for reward functions with bounded support. Similar consideration is given to the optimistic-UCB policy (Lancewicki et al., 2021), and however, in all our experimental settings, optimistic-UCB appears to suffer linear regret under our considered time horizon, and hence we omitted it in the results.

We tested the algorithms with different reward distributions. We consider Poisson distributed arms and exponential reward distributions to represent general reward settings. We use Binomial distributions to represent bounded reward setting. The final consideration is given to Gaussian reward distributions.

Experiment 1: Poisson reward distributions
$K = 5$, $\vec{\lambda} = (1.1, 1.0, 0.5, 0.3, 0.1)$

Experiment 2: Poisson reward distributions
$K = 3$, $\vec{\lambda} = (1.1, 0.9, 0.5)$

Experiment 3: Gaussian reward distributions
$K = 3$, $\vec{\mu} = (4, 3.95, 3.1)$, $\vec{\sigma} = (1, 1, 1)$

Experiment 4: Gaussian reward distributions
$K = 2$, $\vec{\mu} = (4, 3.5, 3.1)$, $\vec{\sigma} = (1, 1.9, 1.3)$

Experiment 5: Binomial reward distributions
$K = 2$, $\vec{n} = (5, 5)$, $\vec{p} = (0.3, 0.5)$

Experiment 6: Binomial reward distributions
$K = 5$, $\vec{n} = (3, 5, 4, 3, 5)$,
$\vec{p} = (0.38, 0.72, 0.63, 0.51, 0.46)$

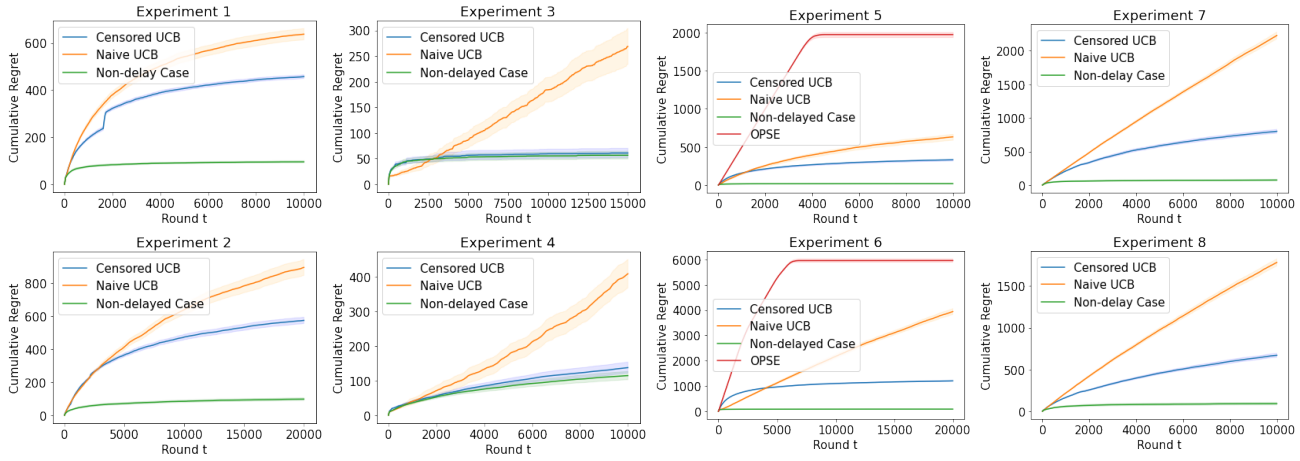

Figure 1: Expected regret under scenarios representing general, bounded and sub-Gaussian reward distributions.

Experiment 7: Exponential reward distributions $K = 3$, $\vec{\lambda} = (0.6, 0.75, 0.9)$

Experiment 8: Exponential reward distributions $K = 3$, $\vec{\lambda} = (0.62, 0.7, 1.25)$

Under Experiment 1 and Experiment 2, we consider unknown horizon T with $T_0 = 1$, $a, b = 2$ for the doubling trick, and $m(t) = \log\log(T)$. Censored UCB yields a regret of $\mathcal{O}(\log(T)\log\log(T)^2)$, which matches our theoretical results. The gap compared to $\mathcal{O}(\log(T))$ under non-delayed case is incurred due to the substantial delays caused by right-skewed long tails of Poisson distributions. Similar results are observed in Experiments 7 and 8 under the exponential reward setting which is usually adopted by survival analysis in clinical trials.

Experiment 3 and Experiment 4 shows that the regret of our algorithm is very close to that of classic UCB without delays. The regret of naive UCB seems to grow linearly in our experiments with Gaussian rewards, bounded rewards and exponential rewards, indicating that under reward-dependent delays, the *observed* rewards can give a biased estimation of the true rewards as smaller reward will be observed earlier. When a substantial portion of rewards remain unobserved, the expected observed reward of the best arm might be smaller than that of a sub-optimal arm. Finally, in $[0, M]$ bounded rewards, OPSE uses M to construct the UCB for all rewards that are not observed. When large amount of rewards are missing, this estimate does not provide much information on the actual distributions. In comparison, our algorithm uses censored reward that gives a more accurate estimation, yielding a much better performance than OPSE. Experiment 5 and Experiment 6 also show that the regret of our algorithm is of the same order as that of classic UCB without delays.

7 CONCLUSION

We have studied multi-armed bandits with (nearly) perfect reward-dependent delays. While the problem formulation is motivated from clinical settings where the delay in getting a response depends on the effectiveness of the treatment, the framework can be used to model other time-to-event data, such as initial breakthrough postoperative pain and/or failure of an implanted medical device. It can also be applied to other domains such as customer churns, product failure, social sciences (e.g. duration of marriage/employment), and epidemiology (e.g. time to infection). Reward-dependent delay is mostly unaddressed in the literature and is more challenging since the observed rewards lead to a biased estimation of the true reward distributions. Under general reward distributions, we present algorithms that achieve near-optimal regret, with tighter regret guarantees under common assumptions on the reward distribution. We also present problem-independent regret bounds. For Gaussian rewards, our algorithm matches the worst case regret (under reward-independent delays) of Joulani et al. (2013) up to a logarithmic factor, and the logarithmic factors can be further removed for bounded rewards.

References

Agrawal, S. and Goyal, N. (2017). Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24.

Arya, S. and Yang, Y. (2020). Randomized allocation with nonparametric estimation for contextual multi-armed bandits with delayed rewards. *Statistics & Probability Letters*, 164:108818.

Atan, O., Zame, W. R., and Schaar, M. (2019).

Sequential patient recruitment and allocation for adaptive clinical trials. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1891–1900. PMLR.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

Aziz, M., Kaufmann, E., and Riviere, M.-K. (2021). On multi-armed bandit designs for dose-finding clinical trials. *The Journal of Machine Learning Research*, 22(1):686–723.

Besson, L. and Kaufmann, E. (2018). What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*.

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

Chick, S., Forster, M., and Pertile, P. (2017). A bayesian decision theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1439–1462.

Chick, S. E., Gans, N., and Yapar, Ö. (2022). Bayesian sequential learning for clinical trials of multiple correlated medical interventions. *Management Science*, 68(7):4919–4938.

Degenne, R. and Perchet, V. (2016). Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 1587–1595. PMLR.

Driscoll, J. J. and Rixe, O. (2009). Overall survival: still the gold standard: why overall survival remains the definitive end point in cancer clinical trials. *The Cancer Journal*, 15(5):401–405.

Dudík, M., Hsu, D. J., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*.

Eick, S. G. (1988a). Gittins procedures for bandits with delayed responses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 50(1):125–132.

Eick, S. G. (1988b). The two-armed bandit with delayed responses. *The Annals of Statistics*, pages 254–264.

Gael, M. A., Vernade, C., Carpentier, A., and Valko, M. (2020). Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR.

Gyorgy, A. and Joulani, P. (2021). Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, pages 3988–3997. PMLR.

Hari, P., Romanus, D., Palumbo, A., Luptakova, K., Rifkin, R. M., Tran, L. M., Raju, A., Farrelly, E., Noga, S. J., Blazer, M., et al. (2018). Prolonged duration of therapy is associated with improved survival in patients treated for relapsed/refractory multiple myeloma in routine clinical care in the united states. *Clinical Lymphoma Myeloma and Leukemia*, 18(2):152–160.

Joulani, P., Gyorgy, A., and Szepesvári, C. (2013). Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR.

Keyvanshokooh, E., Zhalechian, M., Shi, C., Van Oyen, M. P., and Kazemian, P. (2019). Contextual learning with online convex optimization: Theory and application to medical decision-making. *Management Science, to appear*.

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. (2021). Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pages 5969–5978. PMLR.

Langenberg, P. and Srinivasan, R. (1982). On the colton model for clinical trials with delayed observations-dichotomous responses. *Biometrical Journal*, 24(3):287–296.

Mandel, T., Liu, Y.-E., Brunskill, E., and Popović, Z. (2015). The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Marshall, J., Schwartzberg, L. S., Bepler, G., Spetzler, D., El-Deiry, W. S., Xiao, N., Reddy, S. K., Kim, E. S., Poste, G. H., and Raghavan, D. (2016). Novel panomic validation of time to next treatment (tnt) as an effective surrogate outcome measure in 4,729 patients.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., et al. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC medicine*, 16(1):1–15.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. (2018). Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR.

Rosenberger, W. F., Sverdlov, O., and Hu, F. (2012). Adaptive randomization for clinical trials. *Journal of biopharmaceutical statistics*, 22(4):719–736.

Ryeznik, Y., Sverdlov, O., and Wong, W. K. (2012). Doubly adaptive biased coin designs for balancing competing objectives in time-to-event trials. *Statistics and Its Interface*, 5(4):401–413.

Shi, L., Wang, J., and Wu, T. (2023). Statistical inference on multi-armed bandits with delayed feedback. In *International Conference on Machine Learning*.

Shrestha, S. and Jain, S. (2021). A bayesian-bandit adaptive design for n-of-1 clinical trials. *Statistics in Medicine*, 40(7):1825–1844.

Van Der Hoeven, D. and Cesa-Bianchi, N. (2022). Nonstochastic bandits and experts with arm-dependent delays. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Varatharajah, Y. and Berry, B. (2022). A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8):1277.

Vernade, C., Cappé, O., and Perchet, V. (2017). Stochastic Bandit Models for Delayed Conversions. In *Conference on Uncertainty in Artificial Intelligence*.

Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B., and Brueckner, M. (2020). Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR.

Wang, X. (2002). Asymptotic properties of bandit processes with geometric responses. *Statistics & probability letters*, 60(2):211–217.

Williamson, S. F., Jacko, P., and Jaki, T. (2022). Generalisations of a bayesian decision-theoretic randomisation procedure and the impact of delayed responses. *Computational statistics & data analysis*, 174:107407.

Xu, J. and Yin, G. (2014). Two-stage adaptive randomization for delayed response in clinical trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(4):559–578.

Zhang, L. and Rosenberger, W. F. (2007). Response-adaptive randomization for survival trials: the parametric approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 56(2):153–165.

Zhou, Z., Wang, Y., Mamani, H., and Coffey, D. G. (2019a). How do tumor cytogenetics inform cancer treatments? dynamic risk stratification and precision medicine using multi-armed bandits. *SSRN*.

Zhou, Z., Xu, R., and Blanchet, J. (2019b). Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32.

Zimmert, J. and Seldin, Y. (2020). An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pages 3285–3294. PMLR.

Supplementary Material for "Stochastic Multi-Armed Bandits with Strongly Reward-Dependent Delays"

A PRELIMINARIES: AUXILIARY LEMMAS

In this section, we provide some technical lemmas that will be useful in the proofs. First, as the fundamental asymptotic property, we have the following lemma:

Lemma A.1. *If ν is a random variable with $\mathbb{E}|\nu| < \infty$, then we have*

$$\lim_{M\to\infty} \mathbb{E}[\nu \wedge M] = \mathbb{E}[\nu].$$

Proof. We note the fact that for all $M > 0$, $|\nu \wedge M| \leq |\nu|$ holds. So $|\nu|$ is a union integrable bound for the random variables $\nu \wedge M$. Since $\nu \wedge M \to \nu, a.s.$, by dominated convergence theorem,

$$\lim_{M\to\infty} \mathbb{E}[\nu \wedge M] = \mathbb{E}[\nu].$$

□

To obtain the concentric inequalities, we require the following well-known lemma:

Lemma A.2 (Hoeffding's Inequality)**.** *If $X_1, X_2, \cdots, X_n$, are sequence of i.i.d. random variables with mean μ and for every i, $X_i \in [a, b], a.s.$, then, we have, for all $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - n\mu\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

This known result can be found in various tutorials, thus we skip the proof. Finally, we show a property of doubling tricks.

Lemma A.3 (A bound for the constant terms in the exponential doubling trick)**.** *Assume that $T_i = \left\lfloor \frac{T_0}{a} a^{b^i} \right\rfloor$ for $a, b > 1$ and $T_0 \in \mathbb{N}_+$, and $L_T = \min\{i : T_i > T\}$, then, there exist constants $c_1'(a, b, T_0)$ and $c_2'(a, b, T_0)$ such that*

$$L_T \leq c_1'(a, b, T_0) \log\log T,$$

for all $T > c_2'$.

Proof. By definition, we have

$$\frac{T_0}{a} a^{b^{L_T - 1}} \leq T,$$

and this yields

$$L_T \leq 1 + \frac{1}{\log b} \log\left(\frac{\log(aT/T_0)}{\log(a)}\right) \leq \left(2 + \frac{1}{\log b}\right) \log\log T,$$

for $T > c_2'(a, b, T_0) := \max\left\{10, \log\left(\frac{a}{T_0}\right), \exp\left\{e^{\frac{\log(2/\log a)}{\log b}}\right\}\right\}$.

We denote $c_1'(a, b, T_0) = 2 + \frac{1}{\log b}$, this directly implies the result. □

Lemma A.3 is useful to bound the low-order term in the exponential doubling trick processes since we will time a term L_T in the low-order terms.

B OMITTED PROOFS IN SECTION 3

In this section, we give bounds to both the given-horizon setting and the unknown-horizon setting. These bounds finally lead to Theorem 1 and Theorem 2 in the paper. We first show the upper bound of regret of given budget T algorithm Alg_T, then we apply the doubling trick to obtain an upper bound for unknown budget T.

Recall for $c > 1$ and a sub-optimal arm i, we define

$$d_{i,i^*}(c) = 0 \bigvee \inf \left\{ x \in \mathbb{R} : \forall k \geq x, \mathbb{E}[\nu_{i^*} \wedge k] - \mathbb{E}[\nu_i \wedge k] \geq \frac{\Delta_i}{c} \right\},$$

which means the smallest value of N, truncating on which can we lead to an at least $\frac{\Delta_i}{c}$ gap between arm i and the optimal. Such value must exist since for any arm i, $\mathbb{E}[\nu_i \wedge M] \to \mathbb{E}[\nu_i]$ (Lemma A.1). However, if the infimum in the definition of $d_{i,i^*}(c)$ is less than 0, it means that for any $M > 0$, truncated by M, $\mathbb{E}[\nu_{i^*} \wedge M] - \mathbb{E}[\nu_i \wedge M] \geq \frac{\Delta_i}{c}$. Thus in such case, $d_{i,i^*}(c)$ is set to be 0.

We begin with the bounds of the given budget setting. The following theorem gives the upper bound of the expected regret in this case.

Theorem B.1. *Suppose we have a monotone increasing function $m(t)$ such that $m(t) < \frac{t}{2}$ for every $t > 1$, then, for any $c > 1$, the $m(t)$-censored-UCB Algorithm with given total budget T leads to upper bounds of the expected regret*

$$\mathbb{E}[R_T] \leq \sum_{i \neq i^*} \frac{48 m(T)^2 c^2}{\Delta_i} \log(KT) + 2m(T) \sum_{i \neq i^*} \Delta_i + \sum_{i \neq i^*, d_{i,i^*}(c) \neq 0} m^{-1}(d_{i,i^*}(c)) \Delta_i,$$

for any $T \geq 2K$.

Note that in this theorem, there is an increasing term $m(T)$, leading to a greater order than the regret of the classical Multi-armed Bandit problem. When $m(\cdot)$ grows sufficiently slow, although the constant term will be larger, our order will be arbitrarily closed to the classical regret bound. Specifically, if we take $m(x) = \log \log x$, we will get a $\mathcal{O}((\log \log T)^2 \log T)$ order. (However, as explained after Theorem 1, this is not a free lunch.)

In order to prove Theorem B.1, we first present the next concentration inequality.

Proposition B.1 (Restatement of Proposition 1)**.** *For any non-random integer m, integer $1 \leq t \leq T$, and $\delta \in (0, \frac{1}{2})$, we have*

$$\mathbb{P}\left(|\hat{r}_t(m(t), i) - \mathbb{E}[\nu_i \wedge m]| \leq m(t) \sqrt{\frac{2}{N_t(m(t), i)} \log\left(\frac{2T}{\delta}\right)} \right) \geq 1 - \delta.$$

Proof of Proposition B.1. First, by Hoeffding's Inequality, we have

$$\mathbb{P}\left(\left| \frac{X_1 + \cdots + X_N}{N} \right| \leq \sqrt{\frac{2}{N} ln\left(\frac{2}{\delta}\right)} \right) \geq 1 - \delta,$$

which holds for all N with i.i.d. reandom variable sequence $X_1, X_2, \cdots \in [0, 1]$.

Apply the union bound for $N = 1, 2, \cdots, T$, we obtain

$$\mathbb{P}\left(\forall N \in [T] : \left| \frac{X_1 + \cdots + X_N}{N} \right| \leq \sqrt{\frac{2}{N} \log\left(\frac{2T}{\delta}\right)} \right) \geq 1 - \delta.$$

Replacing $X_1, X_2 \cdots$ by the sequence $r_i \wedge M$, and N by $N_t(m(t), i)$, since $N_t(m(t), i) \in [T]$, though it is a random variable, we still have

$$\mathbb{P}\left(|\hat{r}_t(m(t), i) - \mathbb{E}[\nu_i \wedge m]| \leq m \sqrt{\frac{2}{N_t(m(t), i)} \log\left(\frac{2T}{\delta}\right)} \right) \geq 1 - \delta.$$

Now, we take $m = m(t)$ in this theorem and apply the union bound for $0 \leq t \leq T$ and $i \in [K]$. We have, with probability at least $1 - \delta$,

$$|\hat{r}_t(m(t), i) - \mathbb{E}[\nu_i \wedge m(t)]| \leq m(t)\sqrt{\frac{2}{N_t(m(t), i)} ln\left(\frac{2KT^2}{\delta}\right)}, \forall t \in [T], i \in [K]. \tag{1}$$

Thus, we complete the proof of the lemma. □

We are now ready to prove Theorem B.1 with given horizon T.

Proof of Theorem B.1 We first use the concentric inequality presented in Proposition B.1, then by using which we derive the upper bound of the number of pulls on each sub-optimal arm. Finally, we sum the regrets to get a high-probability bound. To get the expectation bound, we select specific δ previously.

Assume the time t is the last time we pull a sub-optimal arm i, this implies that, by the algorithm,

$$\hat{r}_{i^*,t}(m(t)) + m(t)\sqrt{\frac{2}{N_t(m(t), i^*)} \log\left(\frac{2KT^2}{\delta}\right)} \leq \hat{r}_t(m(t), i) + m(t)\sqrt{\frac{2}{N_t(m(t), i)} \log\left(\frac{2KT^2}{\delta}\right)}.$$

This means that with probability at least $1 - \delta$, we have

$$\mathbb{E}[\nu_{i^*} \wedge m(t)] \leq \mathbb{E}[\nu_i \wedge m(t)] + 2m(t)\sqrt{\frac{2}{N_t(m(t), i)} ln\left(\frac{2KT^2}{\delta}\right)}. \tag{2}$$

We then discuss whether the step we are in is big enough to give bounds to $N_t(m(t), i)$.

If $\mathbb{E}[\nu_{i^*} \wedge m(t)] - \mathbb{E}[\nu_i \wedge m(t)] \geq \frac{\Delta_i}{c}$, i.e., $m(t) \geq d_{i,i^*}(c)$, then

$$N_t(m(t), i) \leq \frac{8m(t)^2c^2}{\Delta_i^2} \log\left(\frac{2KT^2}{\delta}\right) \leq \frac{16m(T)^2c^2}{\Delta_i^2} \log\left(\frac{2KT^2}{\delta}\right).$$

Thus, until time t, the total number of times we pull the arm i is bounded by $N_t(m(t), i) + m(t) + 1$. As a result, in this case, the regret contributed by this arm cannot exceed

$$(N_t(m(t), i) + m(t) + 1)\Delta_i.$$

On the other hand, if $\mathbb{E}[\nu_{i^*} \wedge m(t)] - \mathbb{E}[\nu_i \wedge m(t)] < \frac{\Delta_i}{c}$, i.e. $m(t) < d_{i,i^*}(c)$, then the total regret from the arm i is bounded by

$$Reg(i) \leq m^{-1}(d_{i,i^*}(c))\Delta_i.$$

Taking the regret from the initialization steps into account, we obtain that the regret contributed by arm i cannot exceed $m^{-1}(d_{i,i^*}(c))\Delta_i + (N_t(m(t), i) + 2m(t))\Delta_i$.

Combining the results from different i, we have the bound for one-time cumulated regret

$$R_T \leq \sum_{i \neq i^*} \frac{16m(T)^2c^2}{\Delta_i} \log\left(\frac{2KT^2}{\delta}\right) + 2m(T)\sum_{i \neq i^*} \Delta_i + \sum_{i \neq i^*} m^{-1}(d_{i,i^*}(c))\Delta_i,$$

with probability at least $1 - \delta$. Take $\delta = \frac{1}{KT}$, the expected regret

$$\mathbb{E}[R_T] \leq \sum_{i \neq i^*} \frac{48m(T)^2c^2}{\Delta_i} \log(KT) + 2m(T)\sum_{i \neq i^*} \Delta_i + \sum_{i \neq i^*} m^{-1}(d_{i,i^*}(c))\Delta_i,$$

in which we used $T > 2K$. Hence, we complete the proof of Theorem B.1. □

We can begin the proof of the main result of Theorem 1 presented in the manuscript.

Proof of Theorem 1. With the upper bounds we have derived above, we use the exponential doubling trick to extend it to any time T.

We will derive a regret upper bound for any time T. Recall that in the unknown total budget setting, $T_i = \left\lfloor a^{b^i} T_0/a \right\rfloor$, so $T_i - T_{i-1} \leq T_i \leq \frac{T_0}{a} a^{b^i}$ for $i \geq 2$. Assume that $L_T = \min\{i : T_i > T\}$, then we have

$$\begin{aligned}\sum_{j=1}^{L_T} m(T_j - T_{j-1})^2 \log(T_j - T_{j-1}) &\leq m(T_{L_T})^2 \sum_{j=1}^{L_T} \log\left(\frac{T_0}{a} a^{b^j}\right) \\ &= m(T_{L_T})^2 \sum_{j=1}^{L_T} \log(a) b^j + m(T_{L_T})^2 L_T \log\left(\frac{T_0}{a}\right) \\ &= b\log(a) m(T_{L_T})^2 \frac{b^{L_T} - 1}{b-1} + m(T_{L_T})^2 L_T \log\left(\frac{T_0}{a}\right).\end{aligned}$$

By Lemma A.3, there exist constants

$$c_1'(a, b, T_0) := 2 + \frac{1}{\log b}, c_2'(a, b, T_0) := \max\left\{10, \log\left(\frac{a}{T_0}\right), \exp\left\{e^{\frac{\log(2/\log a)}{\log b}}\right\}\right\}, \tag{3}$$

such that for $T > c_2'(a, b, T_0)$, $L_T \leq c_1'(a, b, T_0) \log\log(T)$.

Since $(b^{L_T} - 1)\log(a) \leq b \log\left(a^{b^{L_T - 1}}\right) \leq b\log(T)$, $L_T \leq c_1' \log(\log(T))$ for $T > c_2'(a, b, T_0)$, and further more, $\frac{m(T_{L_T})}{m(T)} \leq \frac{m(T_{L_T})}{m(T_{L_T - 1})} \leq c_0$, we obtain, that for all $T > c_2'(a, b, T_0)$,

$$\begin{aligned}\sum_{j=1}^{L_T} M(T_j - T_{j-1})^2 \log(T_j - T_{j-1}) &\leq \frac{b^2 c_0^2 m(T)^2}{b-1} \log(T) + m(T)^2 c_0^2 c_1'(a, b, T_0) \log\left(\frac{T_0}{a}\right) \log\log(T) \\ &\leq \left(\frac{b^2 c_0^2}{b-1} + c_1'(a, b, T_0) \log\left(\frac{T_0}{a}\right)\right) m(T)^2 \log T.\end{aligned} \tag{4}$$

We denote $C_1(a, b, c_0, T_0) = \frac{b^2 c_0^2}{b-1} + c_1'(a, b, T_0) \log\left(\frac{T_0}{a}\right)$. With those low-order terms in Theorem B.1,

$$\sum_{j=1}^{L_T} m(T_j - T_{j-1}) \sum_{i \neq i^*} \Delta_i \leq c_0 m(T) \sum_{i \neq i^*} \Delta_i L_T \leq \sum_{i \neq i^*} C_2(a, b, c_0, T_0) m(T) \log\log T \Delta_i, \tag{5}$$

in which $C_2(a, b, c_0, T_0) = c_0 c_1'(a, b, T_0)$ and

$$\sum_{j=1}^{L_T} \sum_{i \neq i^*} m^{-1}(d_{i,i^*}(c)) \Delta_i \leq c_1'(a, b, T_0) \log\log T \sum_{i \neq i} m^{-1}(d_{i,i^*}(c)) \Delta_i. \tag{6}$$

Combining Equations (4)(5)(6), the expected regret upper bound is given by

$$\begin{aligned}\mathbb{E}[R_T] \leq \sum_{i \neq i^*} \frac{96 C_1 c^2 m(T)^2}{\Delta_i} \log(T) + C_2(a, b, T_0) m(T) \log\log(T) \sum_{i \neq i^*} \Delta_i \\ + c_1'(a, b, T_0) \log\log(T) \sum_{i \neq i^*} m^{-1}(d_{i,i^*}(c)) \Delta_i + 2\frac{T_0 a^b}{a},\end{aligned}$$

for $T > \max\left\{10, \frac{T_0 a^b}{a}, \log\left(\frac{a}{T_0}\right), \exp\left\{e^{\frac{\log(2/\log a)}{\log b}}\right\}\right\}$, in which $C_1 = \frac{b^2 c_0^2}{b-1} + (2 + \frac{1}{\log b}) \log\left(\frac{T_0}{a}\right)$, $C_2 = c_0(2 + \frac{1}{\log b})$, and $c_1'(a, b, T_0) = (2 + q/\log b)$. This implies the Theorem 1, i.e.,

$$\mathbb{E}[R_T] = \mathcal{O}\left(\sum_{i \neq i^*} \frac{m(T)^2 \log(T)}{\Delta_i}\right).$$

Thus, we complete the proof Theorem 1. □

We next turn to the proof of problem-independent regret bounds (Theorem 2) by first presenting the next proposition.

Proposition B.2. *If, additionally, that $m(\cdot)$ satisfies for all $x > 0$, $\mathbb{E}[m^{-1}(x)(\nu_i - \nu_i \wedge x)] \leq G_1 < \infty$, and a uniform integrability $\mathbb{E}[\nu_i] \leq G_2 < \infty$ then*

$$\sum_{i \neq i^*} m^{-1}(d_{i,i^*}(c))\Delta_i \leq \frac{c}{c-1} K G_1, \quad \sum_{i \neq i^*} \Delta_i \leq K G_2.$$

Proof of Proposition B.2. For $\sum_{i \neq i^*} m^{-1}(d_{i,i^*}(c))\Delta_i$, the bounds are given with assumption that $\mathbb{E}[m^{-1}(x)(\nu_i - \nu_i \wedge x)] < G_1$ uniformly. We may assume $d_{i,i^*}(c) > 0$, otherwise, there is no need to prove. Consider a specific sub-optimal arm i. For $0 < y < d_{i,i^*}(c)$, we write

$$\Delta_i = \mathbb{E}[\nu_{i^*} \wedge y] - \mathbb{E}[\nu_i \wedge y] + \mathbb{E}[\nu_{i^*} - \nu_{i^*} \wedge y] - \mathbb{E}[\nu_i - \nu_i \wedge y].$$

By definition of $d_{i,i^*}(c)$, for any $\epsilon > 0$, there exists $y \in (d_{i,i*}(c) - \epsilon, d_{i,i*}(c))$, such that $\mathbb{E}[\nu_{i^*} \wedge y] - \mathbb{E}[\nu_i \wedge y] < \frac{\Delta_i}{c}$, so

$$\Delta_i = \mathbb{E}[\nu_{i^*} \wedge y] - \mathbb{E}[\nu_i \wedge y] + \mathbb{E}[\nu_{i^*} - \nu_{i^*} \wedge y] - \mathbb{E}[\nu_i - \nu_i \wedge y] \leq \frac{\Delta_i}{c} + \mathbb{E}[\nu_{i^*} - \nu_{i^*} \wedge y] - \mathbb{E}[\nu_i - \nu_i \wedge y].$$

This implies that

$$\left(1 - \frac{1}{c}\right) \Delta_i \leq \mathbb{E}[\nu_{i^*} - \nu_{i^*} \wedge y].$$

We time $m^{-1}(y)$ on both sides of the inequality, obtaining that

$$m^{-1}(y) \left(1 - \frac{1}{c}\right) \Delta_i \leq m^{-1}(y) \mathbb{E}[\nu_{i^*} - \nu_{i^*} \wedge y] \leq G_1.$$

Since $m^{-1}(\cdot)$ is monotone increasing, we obtain that

$$m^{-1}(d_{i,i^*}(c) - \epsilon) \left(1 - \frac{1}{c}\right) \Delta_i \leq G_1, \forall \epsilon > 0.$$

Let $\epsilon \to 0$, we obtain

$$m^{-1}(d_{i,i^*}(c))\Delta_i \leq \frac{c}{c-1} G_1.$$

Finally, the assumption $\mathbb{E}[\nu_i] < G_2$ directly leads to the upper bound of $\sum_{i \neq i^*} \Delta_i$,

$$\sum_{i \neq i^*} \Delta_i \leq K G_2.$$

Hence, we complete the proof of Proposition B.2. □

Proof of Theorem 2. The problem-independent bounds can be derived from Theorem 1 by discussing whether gaps Δ_i are greater or smaller than ϵ, a threshold that is to be discussed. We write the upper bound for the problem-dependent cases as

$$C_1 \sum_{i \neq i^*} \frac{m^2(T) \log(T)}{\Delta_i} + C_2 \sum_{i \neq i^*} m(T) \log\log(T) \Delta_i + \sum_{i \neq i^*} C_3 \log\log(T) c_i + C_4,$$

where C_1, C_2, C_3, C_4 are problem-independent constants specified in the proof of Theorem 1, while c_i depends on the problem, i.e., $c_i = \sum_{i \neq i^*} m^{-1}(d_{i,i^*}(c))\Delta_i$.

For any $\epsilon > 0$, we consider the instance-dependent regret contributed by arm i. If $\Delta_i \leq \epsilon$, then the regret contributed will not exceed $\epsilon N_T(i)$, in which $N_T(i)$ is the total number we pull the arm i when the process is completed. So the regret coming from those arms $\Delta_i \leq \epsilon$ is bounded by ϵT. As a result, the regret bound

$$\begin{aligned}
\mathbb{E}[R_T] &\leq \epsilon T + C_1 \sum_{i \neq i^*, \Delta_i > \epsilon} \frac{m^2(T) \log(T)}{\epsilon} + C_2 \sum_{i \neq i^*} m(T) \log\log(T) \Delta_i + \sum_{i \neq i^*} C_3 \log\log(T) c_i + C_4 \\
&\leq \epsilon T + C_1 \frac{m^2(T) K \log(T)}{\epsilon} + C_2 \sum_{i \neq i^*} m(T) \log\log(T) \Delta_i + \sum_{i \neq i^*} C_3 \log\log(T) c_i + C_4,
\end{aligned}$$

and this holds for any $\epsilon > 0$. Specifically, taking $\epsilon = m(T)\sqrt{\frac{C_1 K \log(T)}{T}}$, we have

$$\mathbb{E}[R_T] \leq 2m(T)\sqrt{C_1 KT\log(T)} + C_2 \sum_{i\neq i^*} m(T)\log\log(T)\Delta_i + \sum_{i\neq i^*} C_3 \log\log(T) c_i + C_4. \tag{7}$$

Next, by Proposition B.2,

$$\sum_{i\neq i^*} \Delta_i \leq \sum_{i\neq i^*} \mathbb{E}[\nu_{i^*}] \leq KG_2 \leq G_2\sqrt{KT}.$$

As a result,

$$C_2 \sum_{i\neq i^*} m(T)\log\log(T)\Delta_i \leq C_2 G_2 m(T)\sqrt{KT}\log\log(T) \leq C_2 G_2 m(T)\sqrt{KT\log(T)},$$

for $T > c_2'(a,b,T_0) + 45$ defined in Eq.(3), where $C_5 = C(a,b,T_0)$ is some constant. Similarly, using Proposition B.2, there exists some constant C'' such that

$$C_3 \sum_{i\neq i^*} \log\log(T) c_i \leq C_3 K G_1 \log\log(T) \leq C_3 G_1 \sqrt{KT\log(T)},$$

for $T > c_2'(a,b,T_0) + 45$, Plugging this into Eq. (7), taking $c = 2$ in Proposition B.2, we have

$$\mathbb{E}[R_T] \leq \tilde{C}_1 m(T)\sqrt{KT\log T} + \tilde{C}_2 = \mathcal{O}(m(T)\sqrt{KT\log T}),$$

for $T > \max\left\{10, \log\left(\frac{a}{T_0}\right), \exp\left\{e^{\frac{\log(2/\log a)}{\log b}}\right\}\right\} + \frac{T_0 a^b}{a}$, where $\tilde{C}_1 = 8\sqrt{b^2 c_0^2/(b-1) + (2 + 1/\log b)\log(T_0/a)} + c_0(2+\log b)G_2 + 2(2+1/\log b)G_1, \tilde{C}_2 = \frac{T_0 a^b}{a} G_2$. As a result, we complete the proof of Theorem 2. □

C OMITTED PROOFS IN SECTION 4.1

In this section, we show the upper bound for a given budget setting and then derive an anytime bound. We begin with the case where the rewards are random variables in interval $[0, M]$ (the setting in Section 4.1). We first have a concentric inequality.

Lemma C.1 (Concentric Inequality for Bounded-M setting)**.** *If $0 \leq r_i \leq M$ holds for all stochastic rewards r_i, integer t satisfies that $1 \leq t \leq T$, and $\delta \in (0, \frac{1}{2})$, we have*

$$\mathbb{P}\left(|\hat{r}_t(M,i) - \mathbb{E}[\nu_i]| \leq M\sqrt{\frac{2}{N_t(M,i)} ln\left(\frac{2T}{\delta}\right)}\right) \geq 1 - \delta.$$

Proof of Lemma C.1. The proof of this lemma is the same as that of proposition B.1 since we notice that $0 \leq r_i/M \leq 1$. Noting that $-1 \leq r_i/M \leq 1$ and they are i.i.d sampled. As a result, by Lemma A.2, we have for any positive integer N,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} r_i - \mathbb{E}[\nu_i]\right| \leq M\sqrt{\frac{2}{N} ln\left(\frac{2}{\delta}\right)}\right) \geq 1 - \delta.$$

Taking union bounds for $1 \leq N \leq T$, noting that $1 \leq N_t(M,i) \leq T$, we have

$$\mathbb{P}\left(|\hat{r}_t(M,i) - \mathbb{E}[\nu_i]| \leq M\sqrt{\frac{2}{N_t(M,i)} ln\left(\frac{2T}{\delta}\right)}\right) \geq 1 - \delta.$$

Thus we complete the proof of Lemma C.1. □

Theorem C.1. *The M-Censored-Bounded UCB algorithm with a given total budget T leads to a regret upper bound*

$$\mathbb{E}[R_T] \leq \frac{64M^2}{\Delta_i}\log(KT) + \max(2M, 4)\sum_{i\neq i^*}\Delta_i.$$

Proof of Theorem C.1. The first K steps of the algorithm will give us at most $\sum_{i\neq i^*}\Delta_i$ regrets. From K to M, since we pull each arm at random, we know that the expected regret during this process is at most $\frac{M}{K}\sum_{i\neq i^*}\Delta_i$. Now consider the last time t the agent pulled a sub-optimal arm i with the algorithm. With probability at least $1-\delta$,

$$\mathbb{E}[\nu_{i^*}] \leq \mathbb{E}[\nu_i] + 2M\sqrt{\frac{2}{N_t(M,i)}\log\left(\frac{2KT^2}{\delta}\right)}.$$

And this will lead to the inequality

$$N_t(M,i) \leq \frac{16M^2}{\Delta_i^2}\log\left(\frac{KT}{\delta}\right).$$

So the total regret contributed by arm i is bounded by

$$Reg(i) \leq \frac{16M^2}{\Delta_i}\log\left(\frac{KT}{\delta}\right) + (M+1)\Delta_i.$$

Summing them up and taking $\delta = \frac{1}{KT}$, will give us the conclusion that

$$\mathbb{E}[R_T] \leq \sum_{i\neq i^*}\frac{64M^2}{\Delta_i}\log(KT) + (M+1+M/K)\sum_{i\neq i^*}\Delta_i.$$

If $M \geq 2$, then $M+1+\frac{M}{K} \leq 2M$, if $M < 2$, then $M+1+\frac{M}{K} \leq 4$, thus completing the proof of Theorem C.1. □

Now we can show the Theorem 3.

Proof of Theorem 3. Given this bound, with similar standard doubling tricks for corresponding problems, we can derive an anytime bound. Using the same deduction of anytime upper bound in the general problem, we deduce the anytime bound. Still, we set that $L_T = \min\{i : T_i > T\}$. Without loss of generality, we assume in each episode, the budget $T_i - T_{i-1}$ is greater than the arms K. Then,

$$\sum_{j=1}^{L_T}\log(T) \leq \sum_{j=1}^{L_T}\log\left(\frac{T_0}{a}a^{b^j}\right) \leq \frac{b^2}{b-1}\log(T),$$

for $T > \frac{T_0 a^b}{a}$. Still, by Lemma A.3,

$$\sum_{j=1}^{L_T} M\sum_{i\neq i^*}\Delta_i \leq \left(2+\frac{1}{\log b}\right)\log\log(T)M\sum_{i\neq i^*}\Delta_i,$$

for any $T > \max\left\{10, \log\left(\frac{a}{T_0}\right), \exp\left\{e^{\frac{\log(2/\log a)}{\log b}}\right\}\right\}$. So we have

$$\mathbb{E}[R_T] \leq \sum_{i\neq i^*}\frac{64M^2b^2}{(b-1)\Delta_i}\log T + 2\left(2+\frac{1}{\log b}\right)\log\log(T)\max(M,2)\sum_{i\neq i^*}\Delta_i,$$

for $T > \max\left\{10, \log\left(\frac{a}{T_0}\right), \exp\left\{e^{\frac{\log(2/\log a)}{\log b}}\right\}\right\} + \frac{T_0 a^b}{a}$. For $M > 2$, the above is directly the Theorem 3. If $M < 2$, we note that the arms are pulled in every episode, so $2\max(M,2)\sum_{i\neq i^*}\Delta_i \leq 4MK \leq 4M\frac{T_0 a^{b^2}}{a}$, as a result, combing them both,

$$\mathbb{E}[R_T] \leq \sum_{i\neq i^*}\frac{64M^2b^2}{(b-1)\Delta_i}\log T + 2\left(2+\frac{1}{\log b}+\frac{T_0 a^{b^2}}{a}\right)\log\log(T)M\sum_{i\neq i^*}\Delta_i.$$

Hence, we complete the proof. □

From Theorem C.1 we can also propose the deduction for problem-independent upper bound.

Proof of Theorem 4. With the standard technique, discussing whether $\Delta_i < \epsilon$ or not, we obtain the bound for the leading term.

For $M \geq 2$ and some $\epsilon > 0$, we consider the problem-dependent regret contributed by arm i. If $\Delta_i \leq \epsilon$, then the regret contributed will not exceed $\epsilon N_T(i)$, in which $N_T(i)$ is the total number we pull the arm i when the process is completed. So the regret coming from those arms $\Delta_i \leq \epsilon$ is bounded by ϵT. For those $\Delta_i > \epsilon$, by Theorem C.1, we have

$$Reg_i(T) \leq \frac{64M^2}{\Delta_i}\log(KT) + 2M\Delta_i \leq \frac{64M^2}{\epsilon}\log(KT) + 2M\Delta_i.$$

Combining these two, the total regret upper bound

$$\begin{aligned}\mathbb{E}[R_T] &\leq \epsilon T + \sum_{i:\Delta_i>\epsilon}\frac{64M^2}{\epsilon}\log(KT) + 2M\sum_{i:\Delta_i>\epsilon}\Delta_i \\ &\leq \epsilon T + \frac{64M^2K}{\epsilon}\log(KT) + 2M\sum_{i\neq i^*}\Delta_i.\end{aligned} \tag{8}$$

The leading term is independent of the problem. So by taking $\epsilon = 8M\sqrt{\frac{K\log(KT)}{T}}$ in Equ.(8), we obtain, for all $T \geq K$, the regret

$$\mathbb{E}[R_T] \leq 16M\sqrt{KT\log(KT)} + 2M\sum_{i\neq i^*}\Delta_i \leq 32M\sqrt{KT\log(T)} + 2KM^2,$$

in which we used the fact that $\Delta_i \leq M$, so $\sum_{i\neq i^*}\Delta_i \leq KM$ and $K < T$.

For those $M < 2$, we observe that $4\sum_{i\neq i^*}\Delta_i \leq 4KM \leq 4M\sqrt{KT}$, so in this case,

$$\mathbb{E}[R_T] \leq 32M\sqrt{KT\log T} + 4M\sqrt{KT} \leq 36M\sqrt{KT\log T}.$$

Hence, we complete the proof of Theorem 4. □

Next, we consider the problem in which $r_i \in [-M_1, M_2]$, where M_1 and M_2 are positive numbers. In this case, the rewards will be observed before M_2. As a result, the censor function $m(t)$ can be selected as constant M_2. In this case, we have the following concentric inequality.

Lemma C.2 (Concentric Inequality for Bounded-M setting)**.** *If $-M_1 \leq r_i \leq M_2$ holds for all stochastic rewards r_i, integer t satisfies that $1 \leq t \leq T$, and $\delta \in (0, \frac{1}{2})$, we have*

$$\mathbb{P}\left(|\hat{r}_t(M_2, i) - \mathbb{E}[\nu_i]| \leq (M_1+M_2)\sqrt{\frac{2}{N_t(M_2,i)}ln\left(\frac{2T}{\delta}\right)}\right) \geq 1-\delta.$$

Proof of Lemma C.2. Noting that $-1 \leq r_i/(M_1+M_2) \leq 1$ and they are i.i.d sampled. As a result, by Lemma A.2, we have for any positive integer N,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} r_i - \mathbb{E}[\nu_i]\right| \leq (M_1+M_2)\sqrt{\frac{2}{N}ln\left(\frac{2}{\delta}\right)}\right) \geq 1-\delta.$$

Taking union bounds for $1 \leq N \leq T$, noting that $1 \leq N_t(M_2, i) \leq T$, we have

$$\mathbb{P}\left(|\hat{r}_t(M_2, i) - \mathbb{E}[\nu_i]| \leq (M_1+M_2)\sqrt{\frac{2}{N_t(M_2,i)}ln\left(\frac{2T}{\delta}\right)}\right) \geq 1-\delta.$$

Thus we complete the proof of Lemma C.2. □

From this, we can show the (M_1, M_2)-Censored-Bounded UCB algorithm with a given total budget T.

Theorem C.2. *The (M_1, M_2)-Censored-Bounded UCB algorithm with a given total budget T leads to a regret upper bound*

$$\mathbb{E}[R_T] \leq \frac{64(M_1+M_2)^2}{\Delta_i}\log(KT) + 2\max(M_2, 2)\sum_{i\neq i^*}\Delta_i.$$

Proof of Theorem C.2. The first K steps of the algorithm will give us at most $\sum_{i\neq i^*}\Delta_i$ regrets. From K to M_2, since we pull each arm at random, we know that the expected regret during this process is at most $\frac{M_2}{K}\sum_{i\neq i^*}\Delta_i$. Now consider the last time t the agent pulled a sub-optimal arm i with the algorithm. With probability at least $1-\delta$,

$$\mathbb{E}[\nu_{i^*}] \leq \mathbb{E}[\nu_i] + 2(M_1+M_2)\sqrt{\frac{2}{N_t(M_2,i)}\log\left(\frac{2KT^2}{\delta}\right)}.$$

And this will lead to the inequality

$$N_t(M_2,i) \leq \frac{16(M_1+M_2)^2}{\Delta_i^2}\log\left(\frac{KT}{\delta}\right).$$

So the total regret contributed by arm i is bounded by

$$Reg(i) \leq \frac{16(M_1+M_2)^2}{\Delta_i}\log\left(\frac{KT}{\delta}\right) + (M_2+1)\Delta_i.$$

Summing them up and taking $\delta = \frac{1}{KT}$, will give us the conclusion that

$$\mathbb{E}[R_T] \leq \sum_{i\neq i^*}\frac{64(M_1+M_2)^2}{\Delta_i}\log(KT) + (M_2+1+M_2/K)\sum_{i\neq i^*}\Delta_i.$$

If $M_2 \geq 2$, then $M_2+1+\frac{M_2}{K} \leq 2M_2$, if $M_2 < 2$, then $M_2+1+\frac{M_2}{K} \leq 4$. As a result, we have

$$\mathbb{E}[R_T] \leq \frac{64(M_1+M_2)^2}{\Delta_i}\log(KT) + 2\max(M_2,2)\sum_{i\neq i^*}\Delta_i.$$

We then complete the proof of Theorem C.2. □

D OMITTED PROOFS IN SECTION 4.2

In this section, we give proofs to Theorem 5 and 6. We first give the regret upper bounds for the given budget T setting. Again, for $c>1$ and a sub-optimal arm i, we define

$$d_{i,i^*}(c) = 0\bigvee\inf\left\{x\in\mathbb{R}:\forall k\geq x, \mathbb{E}[\nu_{i^*}\wedge k]-\mathbb{E}[\nu_i\wedge k]\geq\frac{\Delta_i}{c}\right\}.$$

Theorem D.1. *For any $c>1$, we define a constant C_i for each arm i and censored function $m(t)$ by*

$$C_i = exp\left(\max\left\{d_{i,i^*}(c), 2\mathbb{E}[\nu_{i^*}], \mathbb{E}[\nu_{i^*}]+2, 8\right\}\right)\Delta_i,$$

The algorithm with Upper Confidence Bound $UCB_i^{m(t)}(t,T)$ defined in section 4.2 and censoring function $m(t) = \log t$ with given budget T will lead to an expected regret bound:

$$\mathbb{E}[R_T] \leq \sum_{i\neq i^*}2C_i + 2\sum_{i\neq i^*}\Delta_i\log(T) + \sum_{i\neq i^*}\frac{112c^2}{\Delta_i}\log(T).$$

Remarkably in the theorem, there are three terms. The first term C_i corresponds to the regrets caused by the delays. The second term is attributed to the censored function, $m(t)=\log t$, coming from the time points from $t-\log(t)$ to t, while the third term is the leading term of the algorithm.

To prove this theorem, we first show the concentric inequalities. First, the following lemma shows that a truncation of a sub-Gaussian distribution is nearly sub-Gaussian, which is useful for the deduction of upper bounds in Theorem D.1.

Lemma D.1. *If a random variable X is 1-sub-Gaussian and a non-random number M is chosen such that $M > \max\{2\mathbb{E}X, \mathbb{E}X + 2\}$ and $(M - \mathbb{E}X)^2 \geq 2M$, then for all $\lambda \in \mathbb{R}$*

$$\mathbb{E}\left(e^{\lambda(X\wedge M - \mathbb{E}[X\wedge M])}\right) \leq e^{\lambda^2} e^{e^{-M}}(1 + e^{-M}),$$

where $X \wedge M = \min\{X, M\}$.

Proof of Lemma D.1. Since $M > \max(2\mathbb{E}X, \mathbb{E}X + 2)$, we have

$$\begin{aligned}
\mathbb{E}[X - X\wedge M] &= \int_0^\infty \mathbb{P}\left(X - X\wedge M \geq t\right) dt \\
&= \int_0^\infty \mathbb{P}\left(X \geq t + M\right) dt \\
&= \int_0^\infty \mathbb{P}\left(X - \mathbb{E}X \geq t + M - \mathbb{E}X\right) dt \\
&\leq \int_0^\infty e^{-\frac{(t+M-\mathbb{E}X)^2}{2}} dt \\
&= \int_{M-\mathbb{E}X}^\infty e^{-\frac{t^2}{2}} dt \\
&\leq \int_{M-\mathbb{E}X}^\infty e^{-t} dt \\
&\leq \int_{M/2}^\infty e^{-t} dt = e^{-M/2}.
\end{aligned} \tag{9}$$

Now for $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda(X\wedge M - \mathbb{E}[X\wedge M])}\right) &\leq \mathbb{E}\left(e^{\lambda X - \lambda\mathbb{E}[X\wedge M]}\right) \\
&= \mathbb{E}\left(e^{\lambda(X-\mathbb{E}X)}\right) e^{\lambda(\mathbb{E}X - \mathbb{E}[X\wedge M])} \\
&\leq e^{\frac{\lambda^2}{2}} e^{\frac{\lambda^2}{2} + \frac{(\mathbb{E}X - \mathbb{E}[X\wedge M])^2}{2}} \\
&\leq e^{\lambda^2} e^{e^{-M}} \leq (1 + e^{-M}) e^{\lambda^2} e^{e^{-M}},
\end{aligned}$$

in which we used the fact that $(\mathbb{E}X - \mathbb{E}[X \wedge M])^2/2 \leq e^{-M}$ from Eq.(9).

For $\lambda < 0$,

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda(X\wedge M - \mathbb{E}[X\wedge M])}\right) &= \left(\mathbb{E}[e^{\lambda X} I(X \leq M)] + \mathbb{E}[e^{\lambda M} I(X > M)]\right) e^{-\lambda\mathbb{E}[X\wedge M]} \\
&\leq e^{-\lambda\mathbb{E}[X\wedge M]}\mathbb{E}\left(e^{\lambda X}\right) + e^{-\lambda\mathbb{E}[X\wedge M]}\mathbb{E}\left(e^{\lambda M} I(X > M)\right).
\end{aligned}$$

We denote $p_1 = e^{-\lambda\mathbb{E}[X\wedge M]}\mathbb{E}\left(e^{\lambda X}\right)$ and $p_2 = e^{-\lambda\mathbb{E}[X\wedge M]}\mathbb{E}\left(e^{\lambda M} I(X > M)\right)$. Firstly,

$$p_1 \leq e^{\frac{\lambda^2}{2}} e^{\lambda\mathbb{E}X - \lambda\mathbb{E}[X\wedge M]}.$$

On the other hand, if $(M - \mathbb{E}X)^2 > 2M$,

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda M} I(X > M)\right) &= e^{\lambda M}\mathbb{P}\left(X > M\right) \\
&= e^{\lambda M}\mathbb{P}\left(X - \mathbb{E}X > M - \mathbb{E}X\right) \\
&\leq e^{\lambda M} e^{-\frac{(M-\mathbb{E}X)^2}{2}} \\
&= e^{\lambda\mathbb{E}X} e^{-\frac{(M-\mathbb{E}X)^2}{2} + \lambda(M-\mathbb{E}X)} \\
&\leq e^{\lambda\mathbb{E}X} e^{-\frac{(M-\mathbb{E}X)^2}{2}} \\
&\leq e^{\lambda\mathbb{E}X} e^{-M},
\end{aligned}$$

so

$$p_2 \le e^{\lambda \mathbb{E}X - \lambda \mathbb{E}[X \wedge M]} e^{-M}.$$

As a result,

$$p_1 + p_2 \le e^{\lambda \mathbb{E}X - \lambda \mathbb{E}[X \wedge M]}(e^{\frac{\lambda^2}{2}} + e^{-M}) \le e^{\frac{\lambda^2}{2}} + e^{-M} \le e^{\lambda^2} e^{e^{-M}}(1 + e^{-M}).$$

For $\lambda = 0$, it is obvious. Hence, we complete the proof. $\square$

From this, we can derive the concentric inequality, using the property that $N_t(m(t), i) \le t - m(t)$.

Proposition D.1. *If $m(t) = \lfloor \log t \rfloor$, then with probability at least $1 - \delta$,*

$$|\hat{r}_t(m(t), i) - \mathbb{E}[\nu_i \wedge m(t)]| \le \sqrt{\frac{4}{N_t(m(t), i))} \log\left(\frac{e^2 K T^2}{\delta}\right)}$$

holds for all $m^{-1}(\max\{d_{i,i^*}(c), 2\mathbb{E}[\nu_{i^*}], \mathbb{E}[\nu_{i^*}] + 2, 8\}) < t \le T$ and $i \in [K]$.

Proof. Denote the constant $m^{-1}(\max\{d_{i,i^*}(c), 2\mathbb{E}[\nu_{i^*}], \mathbb{E}[\nu_{i^*}] + 2, 8\})$ by C. From the technical lemma, we can construct the concentric inequality. For independent identically distributed 1-sub-Gaussian random variable X_i, $i = 1, 2, \cdots, n$, we have, for any $\lambda, p \in \mathbb{R}^+$,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} X_i \wedge M - \mathbb{E}[X_i \wedge M] > p\right) &\le \mathbb{E}\left[\frac{e^{\sum_{i=1}^{n} \lambda(X_i \wedge M - \mathbb{E}[X_i \wedge M])}}{e^{\lambda p}}\right] \\ &\le \frac{e^{\lambda^2 n}(1 + e^{-M})^n e^{ne^{-M}}}{e^{\lambda p}}.\end{aligned}$$

Let $p = n\epsilon$, $\lambda = \epsilon/2$, we have

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i \wedge M) - \mathbb{E}[X_i \wedge M] > n\epsilon\right) \le e^{-n\epsilon^2/4}(1 + e^{-M})^n e^{ne^{-M}}.$$

Apply the same for the other side, we obtain that for any n,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i \wedge M) - \mathbb{E}[X \wedge M]\right| > \epsilon\right) \le 2e^{-n\epsilon^2/4}(1 + e^{-M})^n e^{ne^{-M}}.$$

Return to the problem, where $M = m(t)$, $1 \le n \le t - m(t)$. Since $m(t) = \log(t)$, we have

$$(1 + e^{-M})^n \le (1 + \frac{1}{t})^n \le (1 + \frac{1}{t})^t \le e,$$

and

$$ne^{-M} \le \frac{t - M}{t} \le 1.$$

From this, we obtain that if $1 \le n \le t - M$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i \wedge M) - \mathbb{E}[X \wedge M]\right| > \epsilon\right) \le 2e^2 \cdot e^{-n\epsilon^2/4},$$

i.e., whenever $t > C$, $1 \le n \le t - M$ and $M = m(t) = \log(t)$, we always have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i \wedge M) - \mathbb{E}[X \wedge M]\right| \le \sqrt{\frac{4}{n}\log\left(\frac{2e^2}{\delta}\right)}\right) \ge 1 - \delta.$$

So for this t, fixing $M = m(t)$ and applying the union bound for $1 \le n \le t - M$. Since $1 \le N_t(m(t), i) \le t - M$, we have, with probability at least $1 - \delta$,

$$|\hat{r}_t(m(t), i) - \mathbb{E}[\nu_i \wedge m(t)]| \le \sqrt{\frac{4}{N_t(m(t), i)} \log\left(\frac{2e^2 T}{\delta}\right)}.$$

The total union bound will then be given by, with probability at least $1 - \delta$,

$$|\hat{r}_t(m(t), i) - \mathbb{E}[\nu_i \wedge m(t)]| \le \sqrt{\frac{4}{N_t(m(t), i)} \log\left(\frac{2e^2 KT^2}{\delta}\right)},$$

for all $C \le t \le T$ and $i \in [K]$. Hence, we complete the proof of Proposition D.1. □

Now we can turn to the proof of Theorem D.1.

Proof of Theorem D.1. During the process, we consider the last time we pulled the sub-optimal arm $i \neq i^*$, denoting it as t. If t is not big enough such that $t < m^{-1}(d_{i,i^*}(c))$ or $m(t) < \max\{2\mathbb{E}[\nu_{i^*}], \mathbb{E}[\nu_{i^*}] + 2, 8\}$, then the regret which is contributed from this arm

$$Reg_i(t) \le \Delta_i m^{-1}\left(\max\{d_{i,i^*}(c), 2\mathbb{E}[\nu_{i^*}], \mathbb{E}[\nu_{i^*}] + 2, 8\}\right),$$

which is a constant, denoted as C_i.

If t is big enough that $m(t) > \max_{i \neq i^*} \max\{d_{i,i^*}(c), 2\mathbb{E}[\nu_{i^*}], \mathbb{E}[\nu_{i^*}] + 2, 8\}$, then the concentric inequality will hold: with probability $\ge 1 - \delta$, for all $\max_i C_{i,m} \le t \le T$ and $i \in [K]$,

$$|\hat{r}_t(m(t), i) - \mathbb{E}[\nu_i \wedge m(t)]| \le \sqrt{\frac{4}{N_t(m(t), i)} \log\left(\frac{2e^2 KT^2}{\delta}\right)}.$$

That means

$$\mathbb{E}[\nu_{i^*} \wedge m] \le \mathbb{E}[\nu_i \wedge m] + 2\sqrt{\frac{4}{N_t(m(t), i)} \log\left(\frac{2e^2 KT^2}{\delta}\right)}.$$

Consequently,

$$\frac{\Delta_i}{c} \le 2\sqrt{\frac{4}{N_t(m(t), i)} \log\left(\frac{2e^2 KT^2}{\delta}\right)},$$

and thus

$$N_t(m(t), i) \le \frac{16c^2}{\Delta_i^2} \log\left(\frac{2e^2 KT^2}{\delta}\right).$$

In conclusion, the regret contributed by sub-optimal arm i is bounded by

$$Reg_i(t) \le C_i + \Delta_i(m(T) + 1) + \frac{16c^2}{\Delta_i} \log\left(\frac{2e^2 KT^2}{\delta}\right).$$

Summing $Reg_i(t)$ for each sub-optimal arm i, noting that the regret from the initialization step will not exceed $\sum_{i \neq i^*} \Delta_i$, we have, with probability $\ge 1 - \delta$,

$$R_T \le \sum_{i \neq i^*} C_i + 2\sum_{i \neq i^*} \Delta_i m(T) + \sum_{i \neq i^*} \frac{16c^2}{\Delta_i} \log\left(\frac{2e^2 KT^2}{\delta}\right).$$

Note that $m(t) = \lfloor \log t \rfloor$, the upper bound is $O(\log T)$. Still, we take $\delta = \frac{1}{KT}$, we obtain the expectation regret bound

$$\mathbb{E}[R_T] \le \sum_{i \neq i^*} 2e^{d_{i,i^*}(c)} \Delta_i + 2\sum_{i \neq i^*} \Delta_i \log(T) + \sum_{i \neq i^*} \frac{16c^2}{\Delta_i} \log\left(2e^2 K^2 T^3\right). \qquad \square$$

We now give the complete proof for Theorem 5 as follows.

Proof of Theorem 5. With Theorem D.1, again, we give bounds for the regrets of doubling trick. In this case, we note that for given budget T, the regret upper bound

$$\mathbb{E}[R_T] \leq C\left(\Delta_i + \frac{1}{\Delta_i}\right)\log(T) + 2\sum_{i\neq i^*} C_i,$$

in which $C = 128c^2, C_i = e^{d_{i,i^*}(c)}\Delta_i$. Again, we denote $L_T = \min\{i : T_i > T\}$. Then,

$$\sum_{j=1}^{L_T} \log(T_{i+1} - T_i) \leq \frac{b^2}{b-1}\log(T),$$

in which $T > \frac{T_0 a^b}{a}$. By Lemma A.3,

$$\sum_{i=1}^{L_T} 2\sum_{i\neq i^*} C_i \leq \left(4 + \frac{2}{\log b}\right)\log\log(T)\sum_{i\neq i^*} C_i = o(\log(T)),$$

for $T > \max\left\{10, \log\left(\frac{a}{T_0}\right), \exp\left\{e^{\frac{\log(2/\log a)}{\log b}}\right\}\right\}$.

As a result,

$$\mathbb{E}[R_T] \leq \frac{128c^2b^2}{(b-1)}\left(\Delta_i + \frac{1}{\Delta_i}\right)\log T + \left(4 + \frac{2}{\log b}\right)\log\log(T)\sum_{i\neq i^*} e^{d_{i,i^*}(c)}\Delta_i,$$

for $T > \max\left\{10, \log\left(\frac{a}{T_0}\right), \exp\left\{e^{\frac{\log(2/\log a)}{\log b}}\right\}\right\} + \frac{T_0 a^b}{a}$. Hence, we complete the proof of Theorem 5. □

Finally, we provide the proof for Theorem 6 as follows.

Proof of Theorem 6. We show that if we assume there are upper bounds for $\mathbb{E}[\nu_i] \leq G, \forall i \in [K]$ with some constant G, a problem-independent upper bound under known horizon T would be

$$\mathbb{E}[R_T] \leq 16\sqrt{KT\log(KT)} + C_G K\log(T),$$

where $C_G = (5G + 12)e^{2G+8}$.

We first give bounds for $C_i = e^{d_{i,i^*}(c)}\Delta_i$. In order to give bounds for C_i, since $m(t) = \log(t)$, we only need to give bounds $e^{d_{i,i^*}(c)}\Delta_i$. (Because the other terms, $e^{\max(\mathbb{E}[\nu_i^*]+2,\mathbb{E}[\nu_i^*],8)} \leq e^{2G+8}$ is already bounded)

For a fixed arm $i \neq i^*$, let $N_i = d_{i,i^*}(c)$, we thus have

$$\Delta_i = \mathbb{E}[\nu_{i^*} \wedge N_i] - \mathbb{E}[\nu_i \wedge N_i] + \mathbb{E}[\nu_{i^*} - \nu_{i^*} \wedge N_i] - \mathbb{E}[\nu_i - \nu_i \wedge N_i].$$

By definition, $\mathbb{E}[\nu_{i^*} \wedge N_i] - \mathbb{E}[\nu_i \wedge N_i] \leq \frac{\Delta_i}{c}$, so we have

$$e^{N_i}\Delta_i\frac{c-1}{c} \leq e^{N_i}\mathbb{E}[\nu_{i^*} - \nu_{i^*} \wedge N_i]. \tag{10}$$

If $N_i \leq \max\{2G, 8\}$, then $(2G + 4)e^{2G+4}$ is an upper of Eq. (10). If $N_i > \max(2G, 8)$, by the deduction of Lemma D.1 Eq. (9),

$$e^{N_i}\mathbb{E}[\nu_{i^*} - \nu_{i^*} \wedge N_i] \leq 2$$

As a result,

$$e^{N_i}\Delta_i\frac{c-1}{c} \leq 2 + (2G + 4)e^{2G+8}.$$

This implies that

$$e^{d_{i,i^*}(c)}\Delta_i \leq \frac{c}{c-1}(2G + 6)e^{2G+8}.$$

As a result, for all $c > 1$, we have

$$\mathbb{E}[R_T] \leq \frac{c}{c-1}(2G+6)e^{2G+4} + KG\log(T) + \sum_{i \neq i^*} \frac{112c^2}{\Delta_i}\log(T).$$

Specifically, taking $c = 2$, we have $e^{d_{i,i^*}(c)}\Delta_i \leq (4G+12)e^{2G+8}$. Thus, $C_i \leq (5G+12)e^{2G+8}$. Then, whenever $T > 2(K+G)+1$,

$$\mathbb{E}[R_T] \leq 2(5G+12)e^{2G+8}K + KG\log(T) + \sum_{i \neq i^*} \frac{448}{\Delta_i}\log(T),$$

we then denote $C_G = (11G+24)e^{2G+8}$, consequently,

$$\mathbb{E}[R_T] \leq C_G K\log(T) + \sum_{i \neq i^*} \frac{448}{\Delta_i}\log(T).$$

For any $\epsilon > 0$, we consider the problem-dependent regret contributed by arm i. If $\Delta_i \leq \epsilon$, then the regret contributed will not exceed $\epsilon N_T(i)$, in which $N_T(i)$ is the total number we pull the arm i when the process is completed. So the regret coming from those arms $\Delta_i \leq \epsilon$ is bounded by ϵT. For those $\Delta_i > \epsilon$, by the previous argument, we have

$$Reg_i(T) \leq \frac{448}{\Delta_i}\log(T) + C_G K\log(T).$$

Consequently, the total regret is bounded by

$$\begin{aligned}\mathbb{E}[R_T] &\leq \epsilon T + \sum_{i:\Delta_i > \epsilon} \frac{448}{\epsilon}\log(T) + C_G K\log(T) \\ &\leq \epsilon T + \frac{448K}{\epsilon}\log(T) + C_G K\log(T).\end{aligned} \tag{11}$$

So by taking $\epsilon = \sqrt{\frac{448K\log(T)}{T}}$ in Eq.(11), we obtain, for all $T \geq 1$, the regret

$$\mathbb{E}[R_T] \leq 16\sqrt{7KT\log(T)} + C_G K\log(T),$$

where $C_G = (11G+24)e^{2G+8}$. Hence, we complete the proof of Theorem 6. □

E OMITTED PROOFS IN SECTION 5

In this section, we show the lower bound presented in the paper. We first notice the following lemma presented by Lancewicki et al. (2021), which is a variant of Lemma 11 in Kleinberg et al. (2008). This gives a lower bound of MAB problem without delay.

Lemma E.1 (Lancewicki et al. (2021))**.** *Consider an algorithm ALG^{MAB} for MAB problem without delays. And let $\mathcal{I}_{ber}$ be the set of problems with Bernoulli rewards. If the ALG's regret is bounded by CT^{α} over any problem instance we proposed, then there exists $I \in \mathcal{I}_{ber}$, where Δ is the sub-optimal gap, such that the algorithm has at least a regret of*

$$\mathbb{E}[R_T^{ALG}] \geq \Omega\left(\frac{(1-\alpha)T}{\Delta}\right).$$

Proof of Theorem 7. Using Lemma E.1, we only need to prove there exists a problem instance leading to a regret lower bound $\Omega(d_{i,i^*}\Delta)$ in two-arm case, where

$$d_{i,i^*} = \min\left\{j \in \mathbb{N} : \mathbb{E}\left[\nu_i \wedge k\right] < \mathbb{E}\left[\nu_{i^*} \wedge k\right], \forall k \geq j\right\}.$$

We consider the two arms with sub-optimal arm i and optimal arm i^*. Assume that the optimal arm has a distribution

$$\mathbb{P}(\nu_{i^*} = 0) = 1 - p,$$

and

$$\mathbb{P}(\nu_{i^*} = M + \Delta/p) = p,$$

while the sub-optimal arm i has a distribution

$$\mathbb{P}(\nu_i = 0) = 1 - p,$$

and

$$\mathbb{P}(\nu_i = M) = p.$$

Then, $M = d_{i,i^*}$ in this case. For any time t, the policy has to be made according to the history $\mathcal{H}_{<t}$. For $t < M$, the two distributions $\nu|\mathcal{H}_{<t}$ and $\nu_{i^*}|\mathcal{H}_{<t}$ is totally the same. As a result, before time M, any algorithm cannot discern the difference between the two arms. Thus, we must have a $d_{i,i*}\Delta$ regret. Now we have found two instances $I_1 \in \mathcal{I}_{ber} \subset \mathcal{I}$ and $I_2 \in \mathcal{I}$, in which $\mathcal{I}$ is the set of all problem instances, and I_1 leads to a regret lower bound $\Omega\left(\frac{(1-\alpha)T}{\Delta}\right)$ while I_2 leads to a lower bound $\Omega(d_{i,i^*}\Delta)$. Then, then maximum regret lower bound in $\mathcal{I}$ will be at least $\Omega(\frac{(1-\alpha)T}{\Delta} + d_{i,i^*}\Delta)$ and hence we complete the proof. □

References

Kleinberg, R., Niculescu-Mizil, A., and Sharma, Y. (2008). Regret bounds for sleeping experts and bandits. volume 80, pages 425–436.

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. (2021). Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pages 5969–5978. PMLR.