Journal of Machine Learning Research 25 (2024) 1-55 Submitted 11/22; Revised 8/23; Published 2/24

Exploration, Exploitation, and Engagement in Multi-Armed Bandits with Abandonment

Zixian Yang ZIXIAN@UMICH.EDU
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109, USA

Xin Liu* LIUXIN7@SHANGHAITECH.EDU.CN
School of Information Science and Technology
ShanghaiTech University
Shanghai, China

Lei Ying LEIYING@UMICH.EDU
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Tor Lattimore

Abstract

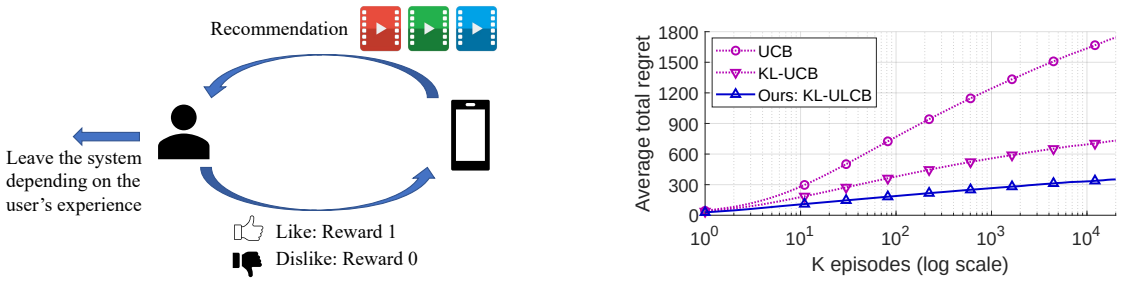
The traditional multi-armed bandit (MAB) model for recommendation systems assumes the user stays in the system for the entire learning horizon. In new online education platforms such as ALEKS or new video recommendation systems such as TikTok, the amount of time a user spends on the app depends on how engaging the recommended contents are. Users may temporarily leave the system if the recommended items cannot engage the users. To understand the exploration, exploitation, and engagement in these systems, we propose a new model, called MAB-A where "A" stands for abandonment and the abandonment probability depends on the current recommended item and the user's past experience (called state). We propose two algorithms, ULCB and KL-ULCB, both of which do more exploration (being optimistic) when the user likes the previous recommended item and less exploration (being pessimistic) when the user does not. We prove that both ULCB and KL-ULCB achieve logarithmic regret, $O(\log K)$, where K is the number of visits (or episodes). Furthermore, the regret bound under KL-ULCB is asymptotically sharp. We also extend the proposed algorithms to the general-state setting. Simulation results show that the proposed algorithms have significantly lower regret than the traditional UCB and KL-UCB, and Q-learning-based algorithms.[1]

Keywords: multi-armed bandit, abandonment, exploration, exploitation, regret bound

1. Introduction

Recommendation algorithms have become increasingly important in many online platforms such as online education, TikTok, YouTube Shorts, advertising platforms, etc. The system interacts with the users to learn their preferences and recommends personalized contents

∗. The work was done when Xin Liu was a postdoctoral research fellow at the University of Michigan.
1. A two-page extended abstract of this paper has appeared at the Allerton Conference in 2022.



(a) A new recommendation system.

(b) Total regret over K episodes.

Figure 1: A new recommendation system and comparison among algorithms.

(learning subjects, videos, songs, products etc.) to each user. These recommendation systems can be modeled as a classic problem called multi-armed bandit (MAB) (Lattimore and Szepesvári, 2020). Each arm in MAB corresponds to a specific type of item in the recommendation system. The recommendation of an item of the ith type is regarded as a pull of arm a_i. Taking recommending short videos as an example, each arm a_i represents *a class of similar videos (e.g. videos from the same dancer, not a single video)*. For simplicity, we assume the reward is 1 if the user likes the recommended item and is 0 otherwise. In this case, it is reasonable to assume that the mean reward of each arm is fixed, which means that the user's preference in different types of items remains unchanged. In a traditional MAB problem, the learner can continue to play the arms with the goal of maximizing the average reward, which either assumes a single user stays in the system for a long period of time or assumes the learner is recommending a single item to each user with a large number of users. While this traditional MAB formulation models recommendation systems such as online advertising well, there are new recommendation systems that are significantly different from these traditional models. In these new recommendation systems, such as TikTok or ALEKS, the learner continuously recommends videos/contents to a user, and the user, other than like or dislike the item, may abandon the system if the recommended items cannot engage the user, and come back later. For example, a user watches TikTok or YouTube Shorts for some period of time, where the duration depends on how interesting/engaging the videos, then leaves the systems, and comes back later, as shown in Figure 1(a).

This makes the problem different from traditional MAB because the objective now is to maximize the total reward per episode (visit) instead of the average reward per pull. Therefore, in addition to finding the most rewarding arm, the learner also needs to continue to engage the user to maximize the number of plays of each episode. Because of the abandonment, the exploration needs to be carefully designed so that the learner should explore (recommend new types of items) when the user is less likely to abandon the system. In other words, we need to consider an exploration-exploitation-engagement tradeoff in this problem. As we can see from Figure 1(b), a well-designed algorithm can significantly outperform the traditional MAB algorithms such as upper confidence bound (UCB) (Auer et al., 2002) and Kullback-Leibler UCB (KL-UCB) (Garivier and Cappé, 2011).

We study this new MAB problem with abandonment, denoted by MAB-A. Consider a recommendation system where the system recommends one item at a time to the user. For example, for mobile phone users, since the screen is small, the system such as a mobile app can

only recommend one item at a time instead of recommending multiple items simultaneously. The user may or may not like the item, and they may abandon the system with a certain probability (called abandonment probability in this paper) based on current and previous experience. The objective is to maximize the total reward per episode, where an episode ends when the user abandons (leaves) the system temporarily.

For traditional MAB problems, classic index policy algorithms such as UCB and KL-UCB work well and KL-UCB achieves the instance-dependent lower bound for traditional MAB with Bernoulli rewards (Garivier and Cappé, 2011; Lai and Robbins, 1985). However, these traditional algorithms are not suitable for the new MAB-A problem, since they do not consider the abandonment and do not use the state information (the user's experience). Hence, they may not be optimal. We propose to use a state-dependent exploration-exploitation mechanism, which does more exploration (being optimistic) when the user is less likely to abandon the system and less exploration (being pessimistic) when the user is more likely to abandon the system. Our algorithms are based on both an upper confidence bound and a lower confidence bound. Our main contributions are as follows:

- **Baseline:** First, we characterize the baseline by showing that a genie-aided optimal policy for MAB-A problem is always pulling the optimal arm (**Lemma 1**).

- **Sharp bounds:** We propose two algorithms based on upper and lower confidence bounds, named as Upper and Lower Confidence Bounds (ULCB) and Kullback-Leibler Upper and Lower Confidence Bounds (KL-ULCB) algorithms. We prove that both algorithms achieve $O(\log K)$ regret bound (**Theorem 2 and Theorem 4**). We further establish an asymptotic lower bound for MAB-A problem and show that KL-ULCB attains the bound (**Theorem 6**), so the regret under KL-ULCB is asymptotically sharp.

- **Extension to a general-state model:** We extend the proposed algorithms to MAB-A problems with a general continuous state space. In particular, we propose four algorithms, DISC-ULCB, DISC-KL-ULCB, CONT-ULCB, and CONT-KL-ULCB. We establish $O(\log K)$ upper bounds for DISC-ULCB and DISC-KL-ULCB (**Theorem 11**) and show that the bound for DISC-KL-ULCB is nearly sharp for large n, where n is the number of discretized bins in the algorithm.

- **Numerical evaluation:** Simulation results in Section 5 and Appendix D confirm our theoretical results and show that our algorithms have significantly lower regret than the traditional UCB and KL-UCB algorithms, and have order-wise lower regret than generic reinforcement learning (RL) algorithms like Q-learning.

- **Technical novelty:** In MAB-A, the episode length follows different distributions under different policies, so the regret analysis based on step-by-step coupling, like in the traditional MAB analysis, does not work. We overcome this difficulty by exploiting the performance difference lemma (Kakade and Langford, 2002; Yang et al., 2021) to couple the rewards by the sum of gap functions along *the sample path that follows our algorithm*. On the other hand, MAB-A can be regarded as a special class of Markov decision process (MDP) problems with a terminal state, and is mostly related to the stochastic shortest path (SSP) problem (Bertsekas and Tsitsiklis, 1991). MAB-A, however, is a stochastic *longest* path problem so the existing algorithms and analysis for SSP do not apply. We take advantage

of the special properties of bandits and abandonment to establish a sharp $O(\log K)$ regret bound, while most regret bounds on SSP are $O(\sqrt{K})$.

1.1 Related Work

We are not aware of any work in the literature with the same setting as the MAB-A problem, but there are a few related works. Schmit and Johari (2018) study a setting with abandonment, where the mean reward is an increasing function of the action. The abandonment occurs when the action is larger than the user's threshold, and thus the algorithms should consider the trade-off between getting high rewards and avoiding losing users. In contrast, in our MAB-A setting, the reward is unknown and a higher reward makes the user less likely to abandon the system. The concept of abandonment also appears in the sequential choice bandit problem (Cao and Sun, 2019) and the departing bandit problem (Ben-Porat et al., 2022). However, the abandonment probabilities in their models do not depend on the past experience of the user. Another work by Wu et al. (2018) studies the exploration-exploitation tradeoff in an opportunistic bandit setting, where the regret of pulling a suboptimal arm varies under different environmental conditions. Our proposed algorithms and proof ideas are partly inspired by the exploration-exploitation intuition in their work (Wu et al., 2018). However, the key difference between the opportunistic bandit setting and our MAB-A setting is that there is no abandonment in the opportunistic bandit setting. Also, the state in the MAB-A setting depends on previous rewards while the load (state) in the opportunistic bandit setting does not. The above differences lead to different algorithms and theoretical results. There are two other works studying user retention, (Sabbeh, 2018) and (Cai et al., 2023). Sabbeh (2018) compared different machine learning techniques to predict the probability that a customer will stay with his service provider or switch to another one. Cai et al. (2023) proposed a novel reinforcement learning algorithm, which can significantly improve user retention. However, there is no theoretical performance guarantee in these works.

Note that the MAB-A problem can be modeled as a special case of stochastic shortest path (SSP) problems (Bertsekas and Tsitsiklis, 1991) with non-positive costs. RL algorithms like Q-learning (Watkins, 1989) and Q-learning with UCB (Yang et al., 2021) might be used for the MAB-A problem but these general algorithms do not make use of the special structures in MAB-A and therefore are too complex and not regret optimal, which is verified in the simulation results in Section 5. Other algorithms (Cohen et al., 2021; Chen et al., 2021; Vial et al., 2021; Tarbouriech et al., 2021) are designed for SSP problems with non-negative costs, which are fundamentally different from MAB-A since MAB-A tries to maximize the episode length but SSP problems with non-negative costs may not. Hence, these algorithms cannot be applied to the MAB-A problem. Besides, only $O(\sqrt{K})$ instead of $O(\log K)$ regret bounds are proved in these papers.

2. Model and Preliminaries

The MAB-A problem is defined as follows. Let M ($M \geqslant 2$) be the number of arms and denote the set of arms by $\{a_1, a_2, \cdots, a_M\}$. Assume that the rewards of pulling the arms are i.i.d. Bernoulli random variables with unknown mean $\mu(a_i)$, $i \in \{1, 2, ..., M\}$. Consider K episodes in total, where each episode represents a single visit of a user and an episode ends when the user abandons the system temporarily. The process of the k^{th} episode goes as follows. At

step $h = 1$, an initial state $S_{k,1} \in \{0, 1\}$ is sampled from an arbitrary distribution. At step $h = 2, 3, ...$, the state is defined by $S_{k,h} := R_{k,h-1}$, where $R_{k,h-1}$ is the reward obtained at the previous step $h - 1$. Then an arm $A_{k,h} \in \{a_1, ..., a_M\}$ is pulled and a Bernoulli random reward $R_{k,h} \in \{0, 1\}$ is obtained with mean $\mu(A_{k,h})$. Given $(S_{k,h}, R_{k,h})$, abandonment occurs with probability $q(S_{k,h}, R_{k,h})$. If the abandonment occurs, the terminal state g is reached, i.e., $S_{k,h+1} = g$, which terminates the current episode k. Otherwise, the process goes to the next step.

Therefore, the process of one episode is an MDP with state space $\mathcal{S} = \{0, 1, g\}$, action space $\mathcal{A} = \{a_1, ..., a_M\}$, and Bernoulli random rewards. The transition graph and details can be found in Appendix A. The state can be interpreted as the experience of the user. At the first step ($h = 1$) in each episode, the initial state $S_{k,1}$ can be interpreted as the user's first impression and is observed by the learner. Note that given $A_{k,h}$, the reward $R_{k,h}$ is independent of $S_{k,h}$. We write $R_{k,h}(A_{k,h})$ when necessary in order to explicitly show the dependency between $R_{k,h}$ and $A_{k,h}$. We will consider a general-state model in Section 4, where the state depends on the rewards received in all previous steps of the current episode. We remark that we first consider the current model, for which we can establish sharp bounds. However, the intuition and exploration strategy obtained from the current model will be applied to the general-state model and nearly sharp bounds can be established based on discretization. We point out that these two models, the simple model and the general-state model cannot capture all the characteristics of real-world systems and hence cannot be directly applied to complex real-world applications, but we discovered that the idea of doing exploration when the user is less likely to abandon the system can help reduce the regret. This intuition could possibly be helpful in the design of low regret algorithms for more complex models such as contextual bandits and be applied in practice when we know when the user is less likely to abandon the system.

We make the following assumption on the problem.

Assumption 1 *Assume* $q(i, j) \geqslant q(i', j')$ *if* $i+j < i'+j'$, $q(0, 0) > 0$, $q(0, 1) < 1$, $q(1, 1) < 1$, *and* $0 < \mu(a_M) \leqslant \mu(a_{M-1}) \leqslant ... \leqslant \mu(a_2) < \mu(a_1) < 1$.

The assumption on $q(\cdot, \cdot)$ implies the abandonment probability becomes larger when the user's experience becomes worse. It also means that the user will continue engaging with the platform when they receive high rewards in hopes that the experience will re-occur (Petrillo, 2021). The assumptions $q(0, 0) > 0$ and $\mu(a_i) < 1$ $\forall i$ ensure that all policies are proper. That is, all policies lead to the terminal state g with probability one, regardless of the initial state (Bertsekas and Tsitsiklis, 1991). Without loss of generality, we let $\mu(a_M) \leqslant \mu(a_{M-1}) \leqslant ... \leqslant \mu(a_2) < \mu(a_1)$. The assumptions $\mu(a_M) > 0$, $q(0, 1) < 1$, and $q(1, 1) < 1$ ensure that there is always a positive proportion of time during which the process is in state 1.

To understand the exploration-exploitation-engagement trade-off of MAB-A defined above, we next define the baseline, i.e. the reward under a genie-aided (model-based) optimal policy, which knows the model perfectly. The result is summarized in Lemma 1, which states that the optimal policy is always pulling arm a_1. The proof can be found in Appendix B.1.

Lemma 1 *Let Assumption 1 hold. The genie-aided optimal policy* π^* *is always pulling arm* a_1.

Let $\pi : \mathcal{S} \times \Phi \to \mathcal{A}$ denote a deterministic policy such that $A_{k,h} = \pi(S_{k,h}, \phi_{k,h})$, where $\phi_{k,h} \in \Phi$ is the historical samples till step h of episode k (not including the current step), i.e.,

$$\phi_{k,h} = (S_{1,1}, A_{1,1}, R_{1,1}, ..., S_{k,1}, A_{k,1}, R_{k,1}, ..., S_{k,h-1}, A_{k,h-1}, R_{k,h-1}).$$

Let $\Pi := \{\pi : \mathcal{S} \times \Phi \to \mathcal{A}\}$ denote the set of all such policies. Let $I_k(\pi, s, \varphi)$ denote the number of steps taken to reach the terminal state g given the current state s and the historical samples $\varphi \in \Phi$ under the policy $\pi \in \Pi$ in episode k. Mathematically, let D be a random set such that $D(\pi, s, \varphi) := \{i : S_{k,h+i} = g, S_{k,h} = s, \phi_{k,h} = \varphi, A_{k,h+j} = \pi(S_{k,h+j}, \phi_{k,h+j}), \forall j = 0, 1, ..., i-1\}$, where $S_{k,h}, \phi_{k,h}, A_{k,h}$ (for all h) are random variables under the process controlled by the policy π. Then

$$I_k(\pi, s, \varphi) := \begin{cases} \min D(\pi, s, \varphi), & \text{if } D(\pi, s, \varphi) \neq \varnothing; \\ \infty, & \text{if } D(\pi, s, \varphi) = \varnothing. \end{cases}$$

Similarly, let $I_k(\pi^*, s)$ denote the number of steps taken to reach the terminal state g given the current state s under π^* in episode k, i.e.,

$$I_k(\pi^*, s) := \begin{cases} \min D^*(s), & \text{if } D^*(s) \neq \varnothing; \\ \infty, & \text{if } D^*(s) = \varnothing, \end{cases}$$

where $D^*(s) := \{i : S_{k,h+i} = g, S_{k,h} = s, A_{k,h+j} = a_1, \forall j = 0, 1, ..., i-1\}$, in which $S_{k,h}, A_{k,h}$ (for all h) are random variables under the process controlled by the policy π^*.[2]

The objective is to find a policy $\pi \in \Pi$ to minimize the expected regret (over K episodes) defined by

$$\mathbb{E}[\mathrm{Reg}_\pi(K)] = \mathbb{E}\left[\sum_{k=1}^{K} \sum_{h=1}^{I_k(\pi^*, S_{k,1})} R_{k,h}(a_1)\right] - \mathbb{E}\left[\sum_{k=1}^{K} \sum_{h=1}^{I_k(\pi, S_{k,1}, \phi_{k,1})} R_{k,h}(\pi(S_{k,h}, \phi_{k,h}))\right]. \quad (1)$$

3. Main Results and the Proof Roadmap

In this section, we first present two algorithms for the MAB-A problem. One is ULCB, which uses an upper or lower confidence bound depending on the state for exploration and exploitation. The other one is KL-ULCB algorithm, which uses KL divergence for the confidence bounds.

3.1 Algorithms

We propose the ULCB algorithm, which is an index policy like UCB algorithm but the difference is that ULCB uses state-dependent indices, as shown in Algorithm 1. Firstly, the ULCB algorithm plays each arm once by Round-Robin. After that, at step h of episode k, if the state $S_{k,h} = 0$, we let

$$\tilde{\mu}_t^0(a) = \bar{\mu}_t(a) + c_0 \sqrt{\frac{\log t + c \log(\log t)}{2N_t(a)}} \quad (2)$$

2. We slightly abuse the notation, not including the policy in the notation $S_{k,h}, A_{k,h}$.

Algorithm 1 ULCB Algorithm

Initialize: $N_1(a) \leftarrow 0$, $\bar{\mu}_1(a) \leftarrow 0$ for all $a \in \mathcal{A}$, $t \leftarrow 1$, c_0, c_1, c.
for episode $k = 1, ..., K$ **do**
  $h \leftarrow 1$, $S_{k,1} \leftarrow$ initial state of episode k, $S_{k,1} \in \{0, 1\}$
  while $S_{k,h} \neq g$ **do**
    if there exists Arm a' such that $N_t(a') = 0$ **then**
      play Arm $A_{k,h} = a'$ and observe $R_{k,h}$ // play each arm once
    else
      if $S_{k,h} = 0$ **then**
        Let $\tilde{\mu}_t^0(a) = \bar{\mu}_t(a) + c_0\sqrt{\frac{\log t + c\log(\log t)}{2N_t(a)}}$ for all $a \in \mathcal{A}$ // indices for state 0
        Take the action $A_{k,h} \in \operatorname{argmax}_a \tilde{\mu}_t^0(a)$ and observe $R_{k,h}$
      else
        Let $\tilde{\mu}_t^1(a) = \bar{\mu}_t(a) + c_1\sqrt{\frac{\log t + c\log(\log t)}{2N_t(a)}}$ for all $a \in \mathcal{A}$ // indices for state 1
        Take the action $A_{k,h} \in \operatorname{argmax}_a \tilde{\mu}_t^1(a)$ and observe $R_{k,h}$
      end if
    end if
    if abandonment occurs **then** $S_{k,h+1} = g$
    else $S_{k,h+1} = R_{k,h}$
    end if
    Define $(S_t, A_t, S_t', R_t) := (S_{k,h}, A_{k,h}, S_{k,h+1}, R_{k,h})$
    /* update $N_{t+1}(a)$ and $\bar{\mu}_{t+1}(a)$ */
    Update: $N_{t+1}(A_t) = N_t(A_t) + 1$ and $N_{t+1}(a) = N_t(a)$ $\forall a \neq A_t$
    Update: $\bar{\mu}_{t+1}(A_t) = \frac{\bar{\mu}_t(A_t)N_t(A_t)+R_t}{N_{t+1}(A_t)}$ and $\bar{\mu}_{t+1}(a) = \bar{\mu}_t(a)$ $\forall a \neq A_t$
    $t \leftarrow t + 1$, $h \leftarrow h + 1$
  end while
end for

for all $a \in \mathcal{A}$, where c and c_0 are constants, t is the time step counting from the first episode, $N_t(a) := \sum_{s=1}^{t-1} \mathbb{1}\{A_s = a\}$ denotes the number of times arm a has been pulled before time step t, and $\bar{\mu}_t(a) := \left(\sum_{s=1}^{t-1} \mathbb{1}\{A_s = a\}R_s\right)/N_t(a)$ denotes the average of rewards of pulling arm a before time step t. Note that we also denote the state, the action, and the reward at time step t by S_t, A_t, and R_t, respectively. Then we take an action $A_{k,h} \in \operatorname{argmax}_a \tilde{\mu}_t^0(a)$. If the state $S_{k,h} = 1$, we let

$$\tilde{\mu}_t^1(a) = \bar{\mu}_t(a) + c_1\sqrt{\frac{\log t + c\log(\log t)}{2N_t(a)}} \tag{3}$$

for all $a \in \mathcal{A}$, where c_1 is a constant. Then we take an action $A_{k,h} \in \operatorname{argmax}_a \tilde{\mu}_t^1(a)$. The algorithm then updates S_{t+1}, $N_{t+1}(a)$, and $\bar{\mu}_{t+1}(a)$. The process goes to the next step or the next episode depending on whether the abandonment occurs or not.

In fact, the indices $\tilde{\mu}_t^0(a)$ and $\tilde{\mu}_t^1(a)$ are the (upper or lower) confidence bounds of the expected reward of arm a. Note that c_0 and c_1 in (2) and (3) are not necessarily positive. Our theoretical results actually indicate that we should use positive coefficient in state 1 and

negative coefficient in state 0, which means optimism (upper confidence bound) in state 1 and pessimism (lower confidence bound) in state 0. This leads to more exploration in state 1 than in state 0.

We also propose the KL-ULCB algorithm, which replaces the indices $\tilde{\mu}_t^0(a)$ and $\tilde{\mu}_t^1(a)$ in (2) and (3) with

$$\tilde{\mu}_t^0(a) = \min\{p : \mathrm{kl}(\bar{\mu}_t(a), p)N_t(a) \leqslant c_0 \log t + c\log(\log t)\} \tag{4}$$

$$\tilde{\mu}_t^1(a) = \max\{p : \mathrm{kl}(\bar{\mu}_t(a), p)N_t(a) \leqslant c_1 \log t + c\log(\log t)\} \tag{5}$$

where $\mathrm{kl}(p_1, p_2)$ is the KL divergence between two Bernoulli random variables with parameters p_1 and p_2. KL-ULCB is similar to ULCB except that KL-ULCB uses KL divergence for the confidence bound instead of directly adding the bonus term. This idea is borrowed from KL-UCB (Garivier and Cappé, 2011). Note that c_0 and c_1 in (4) and (5) are positive. The "min" in (4) and "max" in (5) imply pessimism in state 0 and optimism in state 1.

3.2 Main Results

We next present three theorems, including the regret upper bound on ULCB (**Theorem 2**), the regret upper bound on KL-ULCB (**Theorem 4**), and a regret lower bound (**Theorem 6**) that matches the upper bound of KL-ULCB. We also present the proof idea and roadmap in the next subsection and present the results for the general-state setting in Section 4.

Let $V^*(s)$ and $Q^*(s,a)$ denote the optimal value function and optimal Q-function defined by

$$V^*(s) := \mathbb{E}\left[\sum_{h=1}^{I_k(\pi^*,s)} R_{k,h}(a_1) \middle| S_{k,1} = s\right], \tag{6}$$

$$Q^*(s,a) := \mu(a) + \mathbb{E}\left[\sum_{h=2}^{I_k(\pi^*,S_{k,2})+1} R_{k,h}(a_1) \middle| S_{k,1} = s, A_{k,1} = a\right], \tag{7}$$

for $s \neq g$, and $V^*(g) := Q^*(g,a) := 0$, for any $a \in \mathcal{A}$.

Theorem 2 (Upper bound for ULCB) *Let Assumption 1 hold. Suppose for any $a \neq a_1$,*

$$V^*(0) - Q^*(0,a) \geqslant V^*(1) - Q^*(1,a). \tag{8}$$

Then under ULCB algorithm with $c_0 = -1$, $c_1 = 1$ and $c = 4$, we have

$$\limsup_{K\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_\pi(K)]}{\log K} \leqslant \sum_{i\neq 1} \frac{V^*(1) - Q^*(1,a_i)}{2(\mu(a_1) - \mu(a_i))^2}.$$

The condition (8) means that a suboptimal pull induces more regret (loss) in state 0 than in state 1. This motivates us to do more exploration in state 1 and to be conservative in state 0, i.e., $c_1 > c_0$. With $c_1 = 1$ and $c_0 = -1$, Theorem 2 provides an asymptotic logarithmic upper bound with an instance-dependent constant $\sum_{i\neq 1} \frac{V^*(1)-Q^*(1,a_i)}{2(\mu(a_1)-\mu(a_i))^2}$. We will show later in Section 3.3.1 that the constant term in the upper bound for the traditional

UCB algorithm ($c_0 = c_1 = 1$) could be $\sum_{i\neq 1} \frac{V^*(0)-Q^*(0,a_i)}{2(\mu(a_1)-\mu(a_i))^2}$, which is greater than the one obtained by the ULCB algorithm. In fact, $V^*(1) - Q^*(1, a_i)$ could be significantly smaller than $V^*(0) - Q^*(0, a_i)$ in some cases. Consider a simple example $q(0, 0) = 1$ and $q(0, 1) = q(1, 0) = q(1, 1) = 0$. Then we have $V^*(1) - Q^*(1, a_i) = \frac{\mu(a_1)-\mu(a_i)}{1-\mu(a_1)}$ and $V^*(0) - Q^*(0, a_i) = \frac{\mu(a_1)-\mu(a_i)}{(1-\mu(a_1))^2}$, and thus the upper bound obtained by ULCB algorithm will be significantly better especially when $\mu(a_1)$ is close to 1.

Condition (8) is in terms of the value function and Q-function, which may not be straightforward to verify. Lemma 3 provides a sufficient condition for (8) in terms of $q(i, j)$. See Appendix B.2 for the proof.

Lemma 3 *Let Assumption 1 hold. Assume $q(1, 0) \neq q(0, 0)$ and*

$$\frac{q(0,1) - q(1,1)}{q(0,0) - q(1,0)} \leqslant \min\left\{\frac{1 - q(0,1)}{1 - q(1,1)}, \frac{1 - q(0,0)}{1 - q(1,0)}\right\}. \tag{9}$$

Then for any $a \neq a_1$, we have

$$V^*(0) - Q^*(0, a) \geqslant V^*(1) - Q^*(1, a).$$

One example of the condition (9) is that the difference between $q(0, 1)$ and $q(1, 1)$ is small but the difference between $q(0, 0)$ and $q(1, 0)$ is relatively large. This means that when the user obtains a reward 1, they are more likely to forget their previous reward compared with obtaining a reward 0 when they make the abandonment decision. Hence, intuitively, we should be more conservative in state 0 so that we are more likely to obtain a reward 1 in order to encourage the user to stay in the system. That is the intuition of using optimistic estimate in state 1 and pessimistic estimate in state 0 in the ULCB algorithm.

We show in (29) in Appendix B.2 that at least one of the two cases, condition (8) or $V^*(0) - Q^*(0, a) \leqslant V^*(1) - Q^*(1, a)$ for all a, holds. When the condition (8) does not hold and hence a suboptimal pull induces more regret in state 1 than in state 0, the learner needs to be optimistic in state 0 and pessimistic in state 1. Modified ULCB and KL-ULCB can guarantee the same regret bounds (see Theorems 19, 20, and 21 in Appendix E).

Theorem 4 (Upper bound for KL-ULCB) *Let all the assumptions in Theorem 2 hold. Then using the KL-ULCB algorithm with $c_0 = c_1 = 1$, and $c = 4$, we have*

$$\limsup_{K\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_\pi(K)]}{\log K} \leqslant \sum_{i\neq 1} \frac{V^*(1) - Q^*(1, a_i)}{\mathrm{kl}(\mu(a_i), \mu(a_1))}. \tag{10}$$

Theorem 4 gives a regret upper bound for KL-ULCB. Compared with the result in Theorem 2, the bound in Theorem 4 is better since $\mathrm{kl}(\mu(a_i), \mu(a_1)) \geqslant 2(\mu(a_1) - \mu(a_i))^2$ by Pinsker's inequality. This bound is also better than the one obtained by KL-UCB, $\sum_{i\neq 1} \frac{V^*(0)-Q^*(0,a_i)}{\mathrm{kl}(\mu(a_i),\mu(a_1))}$, which will be illustrated later in Section 3.3.1.

In order to analyze instance-dependent lower bound for MAB-A, similar to the MAB literature (Lai and Robbins, 1985; Lattimore and Szepesvári, 2020), we define the set of all consistent policies by Π_{cons}:

Definition 5 *A policy $\pi \in \Pi$ is consistent, i.e., $\pi \in \Pi_{\text{cons}}$, if for any $\mu(a_1), ..., \mu(a_M)$, $q(0,0)$, $q(0,1)$, $q(1,0)$, $q(1,1)$, and any $\alpha > 0$, $\lim_{K\to\infty} \mathbb{E}[\text{Reg}_\pi(K)]/K^\alpha = 0$.*

Theorem 6 gives an asymptotic lower bound among policies in Π_{cons} for the MAB-A problem.

Theorem 6 (Lower bound) *Let all the assumptions in Theorem 2 hold. For any $\pi \in \Pi_{\text{cons}}$ and any $\mu(a_1)$,...,$\mu(a_M)$, $q(0,0)$, $q(0,1)$, $q(1,0)$, $q(1,1)$ satisfying the assumptions, we have*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\text{Reg}_\pi(K)]}{\log K} \geqslant \sum_{i\neq 1} \frac{V^*(1) - Q^*(1,a_i)}{\text{kl}(\mu(a_i), \mu(a_1))}. \tag{11}$$

By Theorem 4 and Theorem 6, the regret upper bound obtained by the KL-ULCB algorithm attains the lower bound asymptotically.

Intuitively, the regret upper bound of ULCB (or KL-ULCB) is better than that of UCB (or KL-UCB) because UCB (or KL-UCB) will explore in state 0 where the risk of abandonment is higher and hence the instantaneous regret of a wrong decision in state 0 is higher than that in state 1.

3.3 Proof Roadmap

In this section, we present the proof roadmap of Theorems 2, 4, and 6. Before that, we first illustrate the main intuition behind the proofs.

The first challenge in the proofs is how to couple the rewards from two policies, i.e., the two terms in the regret definition (1). Note that we cannot subtract the rewards step by step like the traditional proof for MAB, since the episode lengths I_k for the two different policies are two different random variables. We use the performance difference lemma (Kakade and Langford, 2002; Yang et al., 2021) in the RL literature to couple the rewards by the sum of gap functions $(V^*(s) - Q^*(s,a))$ along the sample path that follows the policy of the algorithm. The gap function represents the regret induced by pulling a suboptimal arm a in state s assuming all the future actions follow the optimal policy. Then we can deal with the regret step by step and further decompose the regret of each step into state 1 and state 0.

In the proof for the upper bound for ULCB (Theorem 2), we managed to bound the regret induced in state 0 by a constant. First, since there is always a positive proportion of time during which the process is in state 1 and optimistic estimates are used in state 1, we can show that the number of optimal pulls in state 1 scales linearly with t with high probability (Lemma 7). The intuition is that optimistic estimates encourage exploration, which induces only logarithmic number of suboptimal pulls. Hence, the confidence intervals around the optimal arm a_1 should be tight enough. In state 0, it can be proved that *under the pessimistic estimation,* the estimate of a suboptimal arm $\tilde{\mu}_t^0(a)$ is less than the true mean $\mu(a)$ with high probability. Since $\mu(a) \leqslant \mu(a_1)$, $\tilde{\mu}_t^0(a)$ will also be less than $\tilde{\mu}_t^0(a_1)$ with high probability by the tightness of $\tilde{\mu}_t^0(a_1)$ (a result from the analysis in state 1). Hence, a_1 will be pulled with high probability in state 0, which implies a constant upper bound of suboptimal pulls in state 0 (Lemma 8). Then it remains to bound the number of suboptimal pulls in state 1, which can be bounded using the techniques in (Garivier and Cappé, 2011) and an upper bound of episode length (Lemma 9). The proof for the upper bound for KL-ULCB (Theorem 4) is similar to that for ULCB. The difference is that we use concentration inequalities for the KL divergence.

For the lower bound (Theorem 6), we first bound the regret below by the number of suboptimal pulls multiplied by the gap function in state 1 since the gap function in state 1 is smaller than that in state 0. Then it remains to bound the number of suboptimal pulls below. The idea is similar to the proof of the instance-dependent lower bound for the MAB problem (Lai and Robbins, 1985), but the difference is that the horizon (total number of pulls) in MAB-A is not a constant but a random variable. Our idea is to use a simple lower bound, i.e., the horizon is greater than the number of episodes.

Our upper bound matches the lower bound thanks to the regret decomposition, the constant regret in state 0, and the sharpness of the bound in state 1 using KL divergence. Next, we will present the regret decomposition and the proof for Theorem 2 in more details. See Appendix B.7 and B.8 for the proofs of Theorem 4 and Theorem 6, respectively.

3.3.1 Regret Decomposition

Our results and the proofs start from the regret decomposition.

We first define value function and Q function to facilitate the analysis. Define

$$V^{\pi}(s,\varphi) := \mathbb{E}\left[\sum_{h=1}^{I_k(\pi,s,\varphi)} R_{k,h}(\pi(S_{k,h},\phi_{k,h}))\middle| S_{k,1}=s,\phi_{k,1}=\varphi\right], \tag{12}$$

$$Q^{\pi}(s,\varphi,a) := \mu(a) + \mathbb{E}\left[\sum_{h=2}^{I_k(\pi,S_{k,2},\phi_{k,2})+1} R_{k,h}(\pi(S_{k,h},\phi_{k,h}))\middle| S_{k,1}=s,\phi_{k,1}=\varphi,A_{k,1}=a\right],$$

for any $s \neq g$, and $V^{\pi}(g,\varphi) := Q^{\pi}(g,\varphi,a) := 0$, for any $\varphi \in \Phi$, $a \in \mathcal{A}$, and $\pi \in \Pi$.

By the definitions of V^* in (6) and V^{π} in (12), the expected regret defined in (1) is

$$\mathbb{E}[\mathrm{Reg}_{\pi}(K)] = \sum_{k=1}^{K}\Big[\mathbb{E}\left[V^*(S_{k,1})\right] - \mathbb{E}\left[V^{\pi}(S_{k,1},\phi_{k,1})\right]\Big]. \tag{13}$$

By the performance difference formula in the RL literature, we can decompose the regret into the summation of the gaps between value function and Q function in different states shown as (14).

$$\mathbb{E}[\mathrm{Reg}_{\pi}(K)] = \mathbb{E}\Bigg[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{S_t=0,A_t=a_i\}\left[V^*(0)-Q^*(0,a_i)\right] + \mathbb{1}\{S_t=1,A_t=a_i\}\left[V^*(1)-Q^*(1,a_i)\right]\Bigg], \tag{14}$$

where S_t and A_t are the state and action at time step t following the policy π, and $T(K,\pi) := \sum_{k=1}^{K} I_k(\pi,S_{k,1},\phi_{k,1})$ is the number of pulls over K episodes under policy π. The proof details for (14) can be found in Appendix B.3. From the regret decomposition (14), the terms $V^*(0)-Q^*(0,a_i)$ and $V^*(1)-Q^*(1,a_i)$ can be interpreted as the regrets induced by pulling a suboptimal arm a_i in state 0 and 1, respectively. Thus, the key idea of obtaining a lower regret is first determining which of the two terms is smaller and then putting more exploration in that state.

Suppose that we use traditional UCB or KL-UCB algorithm. Both use the same exploration strategy for state 1 and state 0. Thus, from (14), we obtain an upper bound

$$\mathbb{E}[\mathrm{Reg}_\pi(K)] \leqslant \mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{A_t=a_i\}\right][V^*(0)-Q^*(0,a_i)],$$

where the constant term $V^*(0)-Q^*(0,a_i)$ is worse than $V^*(1)-Q^*(1,a_i)$ in Theorem 2 and Theorem 4. Therefore, the use of state-dependent exploration-exploitation mechanism in our algorithms can help us obtain better upper bounds by reducing the expected number of suboptimal pulls in state 0.

3.3.2 Proof of Theorem 2

The proof mainly includes three steps, which correspond to the following three lemmas as we explained in Section 3.3.

Lemma 7 *Let all the assumptions in Theorem 2 hold. Consider the ULCB algorithm with $c_0=-1$, $c_1=1$, and $c=4$. Let $p_{\min}:=\mu(a_M)\min\{1-q(0,1),1-q(1,1)\}$. Let $\eta\in(0,p_{\min})$ and $\gamma\in(0,\mu(a_1)-\mu(a_2))$ be two constants. There exists a constant T_1 such that for any $t\geqslant T_1$,*

$$\begin{aligned}&\mathbb{P}\left(N_t(a_1)\leqslant\frac{(p_{\min}-\eta)(t-1)}{2}\right)\\ \leqslant&\frac{M-1}{2\gamma^2\exp\left(2\gamma^2c_2(t-1)-4\gamma^2\right)}+\frac{c_3}{c_2(t-1)\left[\log\left(c_2(t-1)\right)\right]^2}+\exp\left(-\frac{\eta^2(t-1)}{2}\right),\end{aligned}$$

where c_2, c_3, and T_1 are constants depending only on $p_{\min}$, η, M, γ, $\mu(a_1)$, and $\mu(a_2)$.

Lemma 7 shows that when t is large enough, the number of optimal pulls scales linearly with t with high probability. The key idea of the proof of Lemma 7 is that when $N_t(a_1)$ is small, a_1 will be pulled with high probability. Lemma 7 looks similar to Lemma 2 in the work of Wu et al. (2018) but we have a tighter bound which requires more effort in the proof. We need to use a "peeling trick" (Garivier and Cappé, 2011) instead of directly using the union bound to prove a tighter bound. The proof of this result is based on the optimistic exploration in state 1. See Appendix B.4 for a complete proof. Lemma 7 is essential since we will show that the confidence bound around the optimal arm is tight enough for large t based on this result. Then we can bound the regret induced by pulling suboptimal arms in state 0 by a constant using pessimistic estimate (lower confidence bound), which is shown by Lemma 8:

Lemma 8 *Let all the assumptions in Theorem 2 hold. Consider the ULCB algorithm with $c_0=-1$, $c_1=1$, and $c=4$. The regret induced in state 0 is bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{S_t=0,A_t=a_i\}\left[V^*(0)-Q^*(0,a_i)\right]\right]\leqslant c_4\sum_{i=2}^{M}\left[V^*(0)-Q^*(0,a_i)\right],\quad(15)$$

where c_4 is a constant which depends only on M, $\mu(a_1)$, $\mu(a_2)$, $p_{\min}$, η, and γ.

The proof idea of Lemma 8 is as follows. We first show that $\mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)$ with high probability. And based on Lemma 7 we can show that $\tilde{\mu}_t^0(a_1)$ and $\mu(a_1)$ are close enough for large t. Hence, for large t, we have $\tilde{\mu}_t^0(a_1) \approx \mu(a_1) \geqslant \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)$ with high probability, which implies that a_1 will be pulled in state 0 with high probability. Hence $\mathbb{P}(S_t = 0, A_t = a_i)$ is small enough so that we can bound (15). See Appendix B.5 for a complete proof.

We then bound the regret induced in state 1 by a term of order $\log K$ shown by Lemma 9:

Lemma 9 *Let all the assumptions in Theorem 2 hold. Consider the ULCB algorithm with $c_0 = -1$, $c_1 = 1$, and $c = 4$. For any $\epsilon > 0$, the regret induced in state 1 is bounded by*

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{S_t = 1, A_t = a_i\}\left[V^*(1) - Q^*(1,a_i)\right]\right]\\ \leqslant &\sum_{i\neq 1}\frac{(1+\epsilon)\left[V^*(1) - Q^*(1,a_i)\right]}{2(\mu(a_1) - \mu(a_i))^2}\log K + o(\log K).\end{aligned}$$

In the proof of Lemma 9, we first use techniques from Garivier and Cappé (2011) since ULCB uses upper confidence bound in state 1, and then we bound the term $\mathbb{E}[\log(T(K,\pi))]$ to get a bound of order $\log K$. See Appendix B.6 for a complete proof.

Combining (14) with Lemma 8 and Lemma 9, Theorem 2 is proved.

4. Extension to a General-State Setting

In this section, we extend our results to the general-state setting. We first present an MAB-A model with continuous state space, and then verify that the optimal policy is still always pulling the optimal arm. Next, we propose two types of algorithms and analyze the regret. We obtain the same form of regret lower bound for the general-state setting. We also obtain regret upper bounds for DISC-ULCB and DISC-KL-ULCB algorithms.

4.1 Model and the Optimal Policy

Define the continuous state space by $\mathcal{S} = [0,1] \cup \{g\}$. Define the state $S_{k,h}$ at step h of episode k as an exponential moving average of previous rewards in episode k, i.e.,

$$S_{k,h} := (1-\theta)S_{k,h-1} + \theta R_{k,h-1}$$

for any $k \geqslant 1$ and $h \geqslant 2$, where $\theta \in (0,1)$ is a constant forgetting factor, which means how much the user forgets their previous experience. $S_{k,1} \in [0,1]$ is sampled from an arbitrary distribution. The abandonment probability at step h of episode k is a function of the next state $S_{k,h+1}$, denoted by $q(S_{k,h+1})$.

Assumption 2 *Assume that $0 < q(s_1) \leqslant q(s_2)$ if $s_1 \geqslant s_2$ for any $s_1, s_2 \in [0,1]$, and $0 < \mu(a_M) \leqslant \mu(a_{M-1}) \leqslant ... \leqslant \mu(a_2) < \mu(a_1)$.*

The assumptions on $q(\cdot)$ is reasonable since the abandonment probability becomes larger when the user's experience becomes worse. The positivity assumption on q ensures that all policies are proper. We let $0 < \mu(a_M) \leqslant \mu(a_{M-1}) \leqslant ... \leqslant \mu(a_2) < \mu(a_1)$ as in Assumption 1.

Define the genie-aided (model-based) optimal policy π^* the same way as in the original setting. Then we have Lemma 10. The proof can be found in Appendix C.2.

Lemma 10 *Let Assumption 2 hold. Then the genie-aided optimal policy π^* is always pulling arm a_1.*

4.2 Algorithms and Regret Analysis

We propose DISC-ULCB and DISC-KL-ULCB algorithms, which first discretize the state space $[0,1]$ into n bins, $[0,\frac{1}{n})$, $[\frac{1}{n},\frac{2}{n})$,...,$[\frac{n-1}{n},1]$, and then use the ULCB or KL-ULCB algorithm, where we view any state in $[\frac{n-1}{n},1]$ as state 1 and any state in the other bins as state 0.

We next analyze the regret of these two algorithms. Following the same way as the regret decomposition in (14), we have

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_\pi(K)] =&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\sum_{m=1}^{n-1}\mathbb{1}\left\{S_t\in\left[\frac{m-1}{n},\frac{m}{n}\right),A_t=a_i\right\}\left[V^*(S_t)-Q^*(S_t,a_i)\right]\right.\\ &\left.+\mathbb{1}\left\{S_t\in\left[\frac{n-1}{n},1\right],A_t=a_i\right\}\left[V^*(S_t)-Q^*(S_t,a_i)\right]\right]\end{aligned}$$

for any integer $n\geqslant 2$, where $V^*(S_t)-Q^*(S_t,a_i)$ can be interpreted as the regret induced by pulling arm a_i in state S_t. We consider the case where $V^*(s_1)-Q^*(s_1,a)\leqslant V^*(s_2)-Q^*(s_2,a)$ for any $a\in\mathcal{A}, s_1,s_2\in[0,1], s_1\geqslant s_2$. This case means that the regret induced by pulling a suboptimal arm increases as the state decreases. Some examples can be found in Appendix C.3. In this case, we can obtain an upper bound

$$\begin{aligned}&\mathbb{E}[\mathrm{Reg}_\pi(K)]\\ \leqslant&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\sum_{m=1}^{n-1}\mathbb{1}\left\{S_t\in\left[\frac{m-1}{n},\frac{m}{n}\right),A_t=a_i\right\}\left[V^*\left(\frac{m-1}{n}\right)-Q^*\left(\frac{m-1}{n},a_i\right)\right]\right.\\ &\left.+\mathbb{1}\left\{S_t\in\left[\frac{n-1}{n},1\right],A_t=a_i\right\}\left[V^*\left(\frac{n-1}{n}\right)-Q^*\left(\frac{n-1}{n},a_i\right)\right]\right] \quad (16)\end{aligned}$$

and a lower bound

$$\mathbb{E}[\mathrm{Reg}_\pi(K)]\geqslant\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{A_t=a_i\}\right]\left[V^*(1)-Q^*(1,a_i)\right]. \quad (17)$$

From (17) we can obtain the same regret lower bound as Theorem 6 by following the same proof. For the upper bounds for DISC-ULCB and DISC-KL-ULCB algorithms, we have the following theorem.

Theorem 11 *Let Assumption 2 hold. Let $n\geqslant 2$ denote the number of bins for DISC-ULCB or DISC-KL-ULCB algorithms. Suppose $q(s)<1\ \forall s\in[\frac{n-1}{n},1]$ and*

$$V^*(s_1)-Q^*(s_1,a)\leqslant V^*(s_2)-Q^*(s_2,a) \quad (18)$$

for any $a \in \mathcal{A}, s_1, s_2 \in [0,1], s_1 \geqslant s_2$. *Then using DISC-ULCB algorithm with* $c_0 = -1$, $c_1 = 1$, *and* $c = 4$, *we have*

$$\limsup_{K\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_\pi(K)]}{\log K} \leqslant \sum_{i\neq 1} \frac{V^*\left(\frac{n-1}{n}\right) - Q^*\left(\frac{n-1}{n}, a_i\right)}{2(\mu(a_1) - \mu(a_i))^2}. \tag{19}$$

Using DISC-KL-ULCB algorithm with $c_0 = c_1 = 1$, *and* $c = 4$, *we have*

$$\limsup_{K\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_\pi(K)]}{\log K} \leqslant \sum_{i\neq 1} \frac{V^*\left(\frac{n-1}{n}\right) - Q^*\left(\frac{n-1}{n}, a_i\right)}{\mathrm{kl}(\mu(a_i), \mu(a_1))}. \tag{20}$$

Theorem 11 shows that we have $O(\log K)$ upper bounds for DISC-ULCB and DISC-KL-ULCB. The proof is by (16) and the same method as the proofs of Theorem 2 and Theorem 4, and therefore is omitted. For DISC-KL-ULCB, the asymptotic upper bound is nearly tight for large n. However, if n is large, there is a very small fraction of time when the state of the system is in $[\frac{n-1}{n}, 1]$. It results in a very slow exploration since most of the exploration happens in $[\frac{n-1}{n}, 1]$. Hence, the regret might be large initially despite the fact that the asymptotic regret upper bound is near optimal. To overcome this, we propose a second type of algorithms, CONT-ULCB and CONT-KL-ULCB. For CONT-ULCB, we use indices $\tilde{\mu}_t^s(a)$ as follows

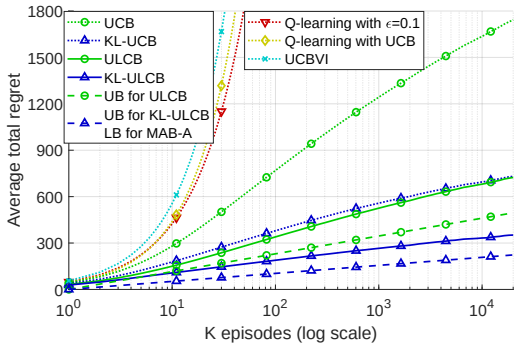
$$\tilde{\mu}_t^s(a) = \bar{\mu}_t(a) + (2s-1)\sqrt{\frac{\log t + c\log(\log t)}{2N_t(a)}}. \tag{21}$$

Similarly, CONT-KL-ULCB uses KL divergence in the indices. $\tilde{\mu}_t^s(a)$ changes gradually from lower confidence bound to upper confidence bound when s changes from 0 to 1, which means that the algorithm changes from exploitation to exploration continuously, which therefore leads to more exploration at the beginning compared to DISC-ULCB and DISC-KL-ULCB.

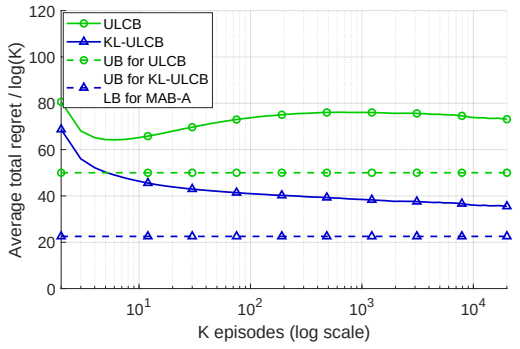
More details, proofs, and simulation results about this extension can be found in Appendix C.

5. Simulation Results

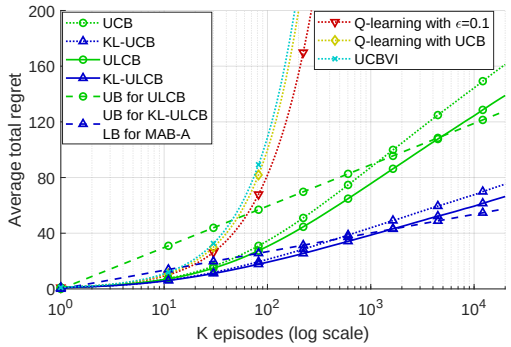
In this section, we present simulation results for the performance of the proposed algorithms. In the simulation, we assume $S_{k,1} = 1$ for simplicity. This is to say that the user assumes a class of items are good if the user has not yet seen the items. Let $M = 2$, $\mu(a_1) = 0.9$, and $\mu(a_2) = 0.8$. Note that for all the algorithms in the simulation, we do not include the log log terms (i.e., $c = 0$) in the indices which are also omitted in (Garivier and Cappé, 2011). We simulated 2×10^4 episodes with 10^7 independent runs. We set $c_1 = 1$, $c_0 = -1$ for ULCB and $c_1 = c_0 = 1$ for KL-ULCB. We also compare our algorithms with Q-learning (Watkins, 1989) with ϵ-greedy, Q-learning with UCB (Yang et al., 2021), and UCBVI (Azar et al., 2017). For Q-learning with ϵ-greedy, at each step, we select a random action with probability ϵ and select a greedy action according to the Q table with probability $1-\epsilon$. At each step, we update the Q table based on the update formula in (Watkins, 1989) with the discount factor $\gamma = 1$. For Q-learning with UCB, at each step, we select a greedy action according to the estimated Q table. At each step, we update the Q table with an additional bonus term



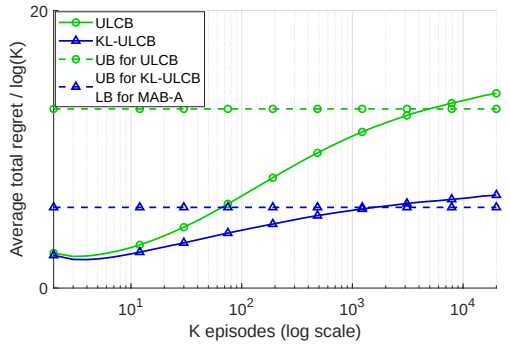
(a) Comparison



(b) UB and LB



(c) Comparison



(d) UB and LB

Figure 2: Simulation results: For (a) and (b), $q(0,0) = 1$, $q(1,1) = q(1,0) = q(0,1) = 0$. For (c) and (d), $q(0,0) = 0.8$, $q(1,0) = q(0,1) = 0.2$, $q(1,1) = 0.1$.

as in (Yang et al., 2021) with the parameter H replaced with the mean episode length of the best policy. For UCBVI, at the beginning of each episode, we first calculate the sample average of the rewards and that of the abandonment probabilities. Then we update the Q table using value iteration as in (Azar et al., 2017) based on the Bellman equation with an additional bonus term. This is a natural model-based RL algorithm for this problem. For Figure 2(a), the 95% confidence bounds are at most ± 8.73. For Figure 2(c), the 95% confidence bounds are at most ± 0.63.

In terms of average cumulative regret, Figure 2(a) and 2(c) show that ULCB outperforms traditional UCB and that KL-ULCB outperforms traditional KL-UCB. The key reason is that ULCB (or KL-ULCB) chooses to explore when the user is less likely to abandon the system, which reduce the risk of abandonment compared to UCB (or KL-UCB). Moreover, our algorithms have order-wise lower regret than Q-learning (Watkins, 1989) with ϵ-greedy, Q-learning with UCB (Yang et al., 2021), and UCBVI (Azar et al., 2017). The reason is that these Q-learning based (model-free or model-based) algorithms have no known regret guarantee for this type of problem, i.e., stochastic longest path problem. It is significantly different from the finite-horizon episodic MDP or discounted MDP problems, where there is either a finite horizon in each episode or a discount factor. From Figure 2(a) and 2(c) we can see that these three algorithms induce a very large regret at the beginning of the learning process, which may be due to severe error propagation during the update of Q values.

Note that the asymptotic upper bound (UB) and lower bound (LB) in Figure 2(a) and 2(c) only consider the $\log K$ term in the regret and ignore the other lower order terms, so only the slopes matter. Figure 2(b) and 2(d) plot the average cumulative regret over K episodes divided by $\log K$. It can be seen that the curves go towards the asymptotic regret upper bound (UB) and the asymptotic lower bound (LB). These results confirm our theoretical results.

See Appendix D for simulation parameters and additional simulation results. Simulations for the general-state setting can be found in Appendix C.5.

6. Conclusion

We studied a new MAB problem with abandonment. The proposed ULCB and KL-ULCB achieve $O(\log K)$ regret, and KL-ULCB is asymptotically sharp. We also extended our algorithms to the general-state setting. Simulation results show that our algorithms outperform UCB, KL-UCB, and Q-learning-based algorithms and confirm our theoretical results about the state-dependent exploration-exploitation mechanism.

Acknowledgments

The work of Zixian Yang and Lei Ying is supported in part by NSF under grants 2002608 and 2001687.

Appendix A. State Transition of the MDP in Section 2

The transition graph of the MDP defined in Section 2 is shown in Figure 3 with state space $\mathcal{S} = \{0, 1, g\}$, action space $\mathcal{A} = \{a_1, ..., a_M\}$, and Bernoulli random rewards.

The transition probabilities $P(s'|s, a)$ while pulling arm a are shown in Table 1. The model can also be extended to the case where users never abandon the system at the first step by defining one more state in which the abandonment probability is 0.

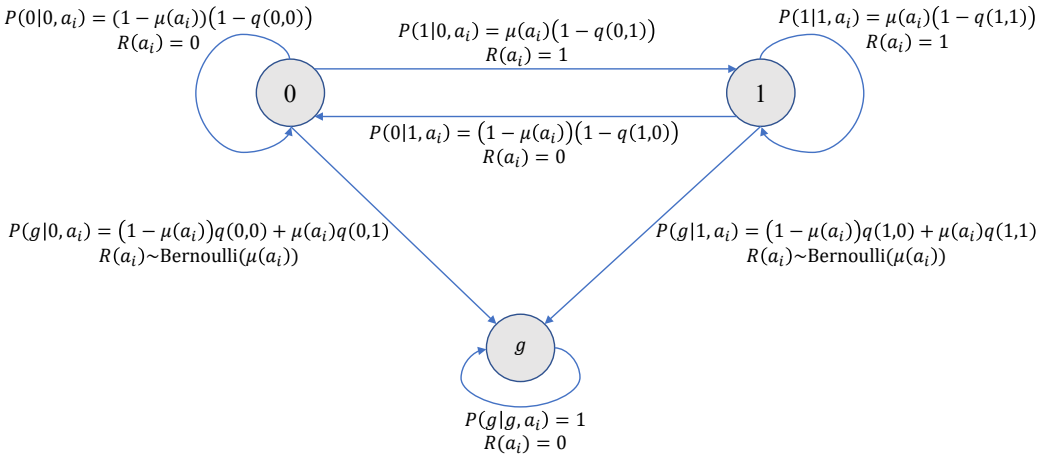


Figure 3: Transition graph for action a_i, $i \in \{1, 2, ..., M\}$.

Appendix B. Missing Proofs

B.1 Proof of Lemma 1: Optimal Policy

If the model is known, this problem can be viewed as a SSP problem (Bertsekas and Tsitsiklis, 1991). Since $\mu(a_i) \leqslant \mu(a_1) < 1, \forall i = 2, ..., M$ and $q(0, 0) > 0$, all policies are proper. Hence, by the results of Bertsekas and Tsitsiklis (1991), there exists a stationary optimal policy. Therefore, it is enough to consider only stationary policies for π^*. Define for any state $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$V^*(s) := \mathbb{E}_{\pi^*}\left[\sum_{h=1}^{\infty} R_{k,h}(A_{k,h}) \mid S_{k,1} = s\right], \tag{22}$$

$P(s'\|s,a)$		Next state s'		
		0	1	g
Current state s	0	$(1-\mu(a))(1-q(0,0))$	$\mu(a)(1-q(0,1))$	$(1-\mu(a))q(0,0) + \mu(a)q(0,1)$
	1	$(1-\mu(a))(1-q(1,0))$	$\mu(a)(1-q(1,1))$	$(1-\mu(a))q(1,0) + \mu(a)q(1,1)$
	g	0	0	1

Table 1: Transition probabilities $P(s'|s, a)$

where $A_{k,h}$ follows the policy π^*, and

$$Q^*(s,a) := \begin{cases} \mu(a) + \mathbb{E}_{\pi^*}\left[\sum_{h=2}^{\infty} R_{k,h}(A_{k,h}) \mid S_{k,1} = s, A_{k,1} = a\right], & s \neq g \\ 0, & s = g \end{cases} \tag{23}$$

where $A_{k,h}, h \geqslant 2$ follows the policy π^*. Note that $V^*(s)$ and $Q^*(s,a)$ do not depend on k since the statistics of the MDPs remain the same among episodes and these MDPs are independent. For $s \neq g$, we have the Bellman equation as follows

$$\begin{aligned} V^*(s) =& \max_a Q^*(s,a) \\ Q^*(s,a) =& \mu(a) + \mathbb{E}\left[\mathbb{E}_{\pi^*}\left[\sum_{h=2}^{\infty} R_{k,h}(A_{k,h}) \,|S_{k,2}\right] \mid S_{k,1} = s, A_{k,1} = a\right] \\ =& \mu(a) + \mathbb{E}\left[V^*(S_{k,2}) | S_{k,1} = s, A_{k,1} = a\right]. \end{aligned} \tag{24}$$

Thus, we have

$$\begin{aligned} Q^*(s,a) =& \mu(a) + P(0|s,a)V^*(0) + P(1|s,a)V^*(1) \\ =& \mu(a) + (1-\mu(a))(1-q(s,0))V^*(0) + \mu(a)(1-q(s,1))V^*(1) \end{aligned} \tag{25}$$

for any $s \in \{0,1\}$ and $a \in \mathcal{A}$. Then we have

$$\begin{aligned} Q^*(1,a) - Q^*(0,a) =& \Big[(1-\mu(a))(1-q(1,0))V^*(0) + \mu(a)(1-q(1,1))V^*(1)\Big] \\ & - \Big[(1-\mu(a))(1-q(0,0))V^*(0) + \mu(a)(1-q(0,1))V^*(1)\Big] \\ =& (1-\mu(a))V^*(0)(q(0,0) - q(1,0)) + \mu(a)V^*(1)(q(0,1) - q(1,1)). \end{aligned} \tag{26}$$

Since by definition we know $V^*(s) \geqslant 0$ for any $s \in \{0,1\}$, and we know $q(0,0) - q(1,0) \geqslant 0$ and $q(0,1) - q(1,1) \geqslant 0$, from the result of (26), we have $Q^*(1,a) - Q^*(0,a) \geqslant 0$ for any $a \in \{a_1, ..., a_M\}$. Therefore, we have

$$\begin{aligned} V^*(1) - V^*(0) =& \max_a Q^*(1,a) - \max_a Q^*(0,a) = \max_a Q^*(1,a) - Q^*(0,a') \\ \geqslant& Q^*(1,a') - Q^*(0,a') \geqslant 0, \end{aligned}$$

where $a' := \text{argmax}_a\, Q^*(0,a)$. Then by (25), for any $i = 2, ..., M$, we have

$$\begin{aligned} Q^*(s,a_1) - Q^*(s,a_i) =& (\mu(a_1) - \mu(a_i)) + (\mu(a_i) - \mu(a_1))(1-q(s,0))V^*(0) \\ & + (\mu(a_1) - \mu(a_i))(1-q(s,1))V^*(1) \\ =& (\mu(a_1) - \mu(a_i)) + (\mu(a_1) - \mu(a_i))\Big[(1-q(s,1))V^*(1) - (1-q(s,0))V^*(0)\Big], \end{aligned} \tag{27}$$

where

$$(1-q(s,1))V^*(1) - (1-q(s,0))V^*(0) \geqslant 0 \tag{28}$$

due to the fact that $V^*(1) \geqslant V^*(0) \geqslant 0$ and $q(s,0) \geqslant q(s,1)$ for any $s \in \{0,1\}$. Therefore, by (27), (28), and $\mu(a_1) \geqslant \mu(a_i), \forall i = 2, ..., M$, we have $Q^*(s,a_1) \geqslant Q^*(s,a_i)$ for any $i = 2, ..., M$ and $s \in \{0,1\}$. Therefore, always pulling Arm a_1 is an optimal policy.

B.2 Proof of Lemma 3

Lemma 3 can be proved by obtaining a lower bound for the ratio $V^*(0)/V^*(1)$. For $i \in \{2, 3, ..., M\}$, we have

$$\begin{aligned}
&[V^*(0) - Q^*(0, a_i)] - [V^*(1) - Q^*(1, a_i)] \\
=&[Q^*(0, a_1) - Q^*(0, a_i)] - [Q^*(1, a_1) - Q^*(1, a_i)] \\
=&[\mu(a_1) - \mu(a_i)]\,[1 + (1 - q(0,1))V^*(1) - (1 - q(0,0))V^*(0)] \\
&- [\mu(a_1) - \mu(a_i)]\,[1 + (1 - q(1,1))V^*(1) - (1 - q(1,0))V^*(0)] \\
=&[\mu(a_1) - \mu(a_i)]\,[(q(0,0) - q(1,0))V^*(0) - (q(0,1) - q(1,1))V^*(1)]\,, \qquad (29)
\end{aligned}$$

where the first and second equalities follow from (24) and Lemma 1. Since $\mu(a_1) > 0$, we have $V^*(1) > 0$. Then by the Bellman equation (24) and Lemma 1, we have

$$\begin{aligned}
&\frac{V^*(0)}{V^*(1)} = \frac{\mu(a_1) + \mu(a_1)(1 - q(0,1))V^*(1) + (1 - \mu(a_1))(1 - q(0,0))V^*(0)}{\mu(a_1) + \mu(a_1)(1 - q(1,1))V^*(1) + (1 - \mu(a_1))(1 - q(1,0))V^*(0)} \\
=&\frac{\mu(a_1) + \frac{(1-q(0,1))}{(1-q(1,1))}\mu(a_1)(1 - q(1,1))V^*(1) + \frac{(1-q(0,0))}{(1-q(1,0))}(1 - \mu(a_1))(1 - q(1,0))V^*(0)}{\mu(a_1) + \mu(a_1)(1 - q(1,1))V^*(1) + (1 - \mu(a_1))(1 - q(1,0))V^*(0)} \\
\geqslant&\frac{\min\left\{\frac{1-q(0,1)}{1-q(1,1)}, \frac{1-q(0,0)}{1-q(1,0)}\right\}[\mu(a_1) + \mu(a_1)(1 - q(1,1))V^*(1) + (1 - \mu(a_1))(1 - q(1,0))V^*(0)]}{\mu(a_1) + \mu(a_1)(1 - q(1,1))V^*(1) + (1 - \mu(a_1))(1 - q(1,0))V^*(0)} \\
=&\min\left\{\frac{1 - q(0,1)}{1 - q(1,1)}, \frac{1 - q(0,0)}{1 - q(1,0)}\right\}, \qquad (30)
\end{aligned}$$

where the inequality is due to the fact that $\min\left\{\frac{1-q(0,1)}{1-q(1,1)}, \frac{1-q(0,0)}{1-q(1,0)}\right\} \leqslant 1$. It follows from (9) and (30) that $\frac{V^*(0)}{V^*(1)} \geqslant \frac{q(0,1)-q(1,1)}{q(0,0)-q(1,0)}$, which implies

$$(q(0,0) - q(1,0))V^*(0) - (q(0,1) - q(1,1))V^*(1) \geqslant 0.$$

Hence, it follows from (29) that $V^*(0) - Q^*(0, a_i) \geqslant V^*(1) - Q^*(1, a_i)$.

We also provide closed-form expressions of the value functions $V^*(1), V^*(0)$ and the Q-functions $Q^*(1, a), Q^*(0, a)$ here. From Lemma 1, we know that the optimal policy is always pulling arm a_1, so $V^*(1) = Q^*(1, a_1)$ and $V^*(0) = Q^*(0, a_1)$. Hence, from the Bellman equation (25), we have the following set of linear equations in terms of $V^*(1)$ and $V^*(0)$

$$\begin{aligned}
V^*(1) =&\mu(a_1) + (1 - \mu(a_1))(1 - q(1,0))V^*(0) + \mu(a_1)(1 - q(1,1))V^*(1) \\
V^*(0) =&\mu(a_1) + (1 - \mu(a_1))(1 - q(0,0))V^*(0) + \mu(a_1)(1 - q(0,1))V^*(1).
\end{aligned}$$

Solving this set of linear equations, we can obtain

$$\begin{aligned}
V^*(1) =&\frac{\mu(a_1)[1 + q(0,0) - \mu(a_1)q(0,0) + \mu(a_1)q(1,0) - q(1,0)]}{\text{denom}} \\
V^*(0) =&\frac{[1 - \mu(a_1)(1 - q(1,1))]V^*(1) - \mu(a_1)}{(1 - \mu(a_1))(1 - q(1,0))}.
\end{aligned}$$

where

$$\begin{aligned}\text{denom} =& q(0,0) + \mu(a_1)[-2q(0,0) + q(1,0) + q(0,1) + q(1,1)q(0,0) - q(1,0)q(0,1)] \\ &+ \mu(a_1)^2[q(0,0) + q(1,1) - q(1,0) - q(0,1) - q(1,1)q(0,0) + q(1,0)q(0,1)].\end{aligned}$$

Substituting the above results of $V^*(1)$ and $V^*(0)$ into (25), we can obtain the expressions for $Q^*(1,a)$ and $Q^*(0,a)$.

B.3 Proof of the Regret Decomposition (14)

From the definition of V^* and Q^* in (6) and (7) and by Lemma 1, we have the following Bellman equation

$$\begin{aligned}V^*(s) =& \max_a Q^*(s,a) = Q^*(s,a_1) \\ Q^*(s,a) =& \mu(a) + \mathbb{E}\left[V^*(S_{k,2}) | S_{k,1} = s, A_{k,1} = a\right]\end{aligned} \tag{31}$$

for $s \neq g$. Similarly, we have the Bellman equation for V^π and Q^π as follows

$$\begin{aligned}V^\pi(s,\varphi) =& Q^\pi(s,\varphi,\pi(s,\varphi)) \\ Q^\pi(s,\varphi,a) =& \mu(a) + \mathbb{E}\Bigg[\mathbb{E}\Bigg[\sum_{h=2}^{I_k(\pi,S_{k,2},\phi_{k,2})+1} R_{k,h}(\pi(S_{k,h},\phi_{k,h}))\,|S_{k,2},\phi_{k,2}\Bigg] \\ &\qquad\qquad \Bigg| S_{k,1} = s, \phi_{k,1} = \varphi, A_{k,1} = a\Bigg] \\ =& \mu(a) + \mathbb{E}\left[V^\pi(S_{k,2},\phi_{k,2}) \mid S_{k,1} = s, \phi_{k,1} = \varphi, A_{k,1} = a\right]\end{aligned} \tag{32}$$

for $s \neq g$.

From (13), the regret induced in episode k is $\mathbb{E}\left[V^*(S_{k,1})\right] - \mathbb{E}\left[V^\pi(S_{k,1},\phi_{k,1})\right]$, which can be decomposed as follows

$$\begin{aligned}&\mathbb{E}\left[V^*(S_{k,1})\right] - \mathbb{E}\left[V^\pi(S_{k,1},\phi_{k,1})\right] \\ =&\mathbb{E}\left[V^*(S_{k,1}) - Q^*(S_{k,1},A_{k,1})\right] + \mathbb{E}\left[Q^*(S_{k,1},A_{k,1}) - V^\pi(S_{k,1},\phi_{k,1})\right] \\ =&\mathbb{E}\left[V^*(S_{k,1}) - Q^*(S_{k,1},A_{k,1})\right] + \mathbb{E}\left[Q^*(S_{k,1},A_{k,1}) - Q^\pi(S_{k,1},\phi_{k,1},A_{k,1})\right] \\ =&\mathbb{E}\left[V^*(S_{k,1}) - Q^*(S_{k,1},A_{k,1})\right] \\ &+ \mathbb{E}\left[\mathbb{E}\left[V^*(S_{k,2})|S_{k,1},A_{k,1}\right] - \mathbb{E}\left[V^\pi(S_{k,2},\phi_{k,2})|S_{k,1},\phi_{k,1},A_{k,1}\right]\right] \\ =&\mathbb{E}\left[V^*(S_{k,1}) - Q^*(S_{k,1},A_{k,1})\right] + \mathbb{E}\left[V^*(S_{k,2}) - V^\pi(S_{k,2},\phi_{k,2})\right] = ... \\ =&\sum_{h=1}^{\infty} \mathbb{E}\left[V^*(S_{k,h}) - Q^*(S_{k,h},A_{k,h})\right],\end{aligned}$$

where $S_{k,h}$, $\phi_{k,h}$, and $A_{k,h}$ are the states, historical samples, and actions following the policy π, respectively. The second equality is due to the fact that $A_{k,h} = \pi(S_{k,h},\phi_{k,h})$, the third equality follows from the Bellman equations (31) and (32), and the fourth equality is by the tower law. The limit in the result is well-defined since $V^*(S_{k,h}) - Q^*(S_{k,h},A_{k,h}) \geqslant 0$. In fact, this regret decomposition borrows from (Yang et al., 2021), and it can also be viewed

as the performance difference formula (Kakade and Langford, 2002) in the RL literature. Then the regret can be further decomposed into the summation of the gaps between value function and Q function in different states shown as follows

$$\begin{aligned}
&\mathbb{E}[\mathrm{Reg}_\pi(K)] = \sum_{k=1}^{K}\sum_{h=1}^{\infty}\mathbb{E}\left[V^*(S_{k,h}) - Q^*(S_{k,h},A_{k,h})\right]\\
=&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{\infty} V^*(S_{k,h}) - Q^*(S_{k,h},A_{k,h})\right]\\
=&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{I_k(\pi,S_{k,1},\phi_{k,1})} V^*(S_{k,h}) - Q^*(S_{k,h},A_{k,h})\right] = \mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} V^*(S_t) - Q^*(S_t,A_t)\right]\\
=&\mathbb{E}\Bigg[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{S_t=0,A_t=a_i\}\left[V^*(0)-Q^*(0,a_i)\right]\\
&\qquad + \mathbb{1}\{S_t=1,A_t=a_i\}\left[V^*(1)-Q^*(1,a_i)\right]\Bigg],
\end{aligned}$$

where the second equality is by monotone convergence theorem, the third equality is by the definition of $I_k(\pi, S_{k,1}, \phi_{k,1})$, $T(K,\pi) \coloneqq \sum_{k=1}^{K} I_k(\pi, S_{k,1}, \phi_{k,1})$ is the number of pulls over K episodes following the policy π, and the last equality follows from the fact that $V^*(s) - Q^*(s,a_1) = 0$ for any s.

B.4 Proof of Lemma 7

Choose a T_1 such that for any $t \geqslant T_1$,

$$\frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \geqslant 2, \text{ and } \sqrt{\frac{\log t + 4\log(\log t)}{\frac{(p_{\min}-\eta)(t-1)}{(M-1)} - 2}} \leqslant (\mu(a_1)-\mu(a_2)) - \gamma.$$

Let $N_t^1(a)$ be the number of times arm $a \in \mathcal{A}$ was pulled in state 1 before time step t. Then

$$\begin{aligned}
&\mathbb{P}\left(N_t(a_1) \leqslant \frac{(p_{\min}-\eta)(t-1)}{2}\right) \leqslant \mathbb{P}\left(N_t^1(a_1) \leqslant \frac{(p_{\min}-\eta)(t-1)}{2}\right)\\
\leqslant&\mathbb{P}\left(N_t^1(a_1) \leqslant \frac{(p_{\min}-\eta)(t-1)}{2}, \sum_{i=1}^{t-1}\mathbb{1}\{S_i=1\} > (p_{\min}-\eta)(t-1)\right)\\
&+\mathbb{P}\left(\sum_{i=1}^{t-1}\mathbb{1}\{S_i=1\} \leqslant (p_{\min}-\eta)(t-1)\right) \qquad (33)
\end{aligned}$$

where the first inequality follows from the fact that $N_t^1(a) \leqslant N_t(a)$. Next we will show that $\mathbb{P}(\sum_{i=1}^{t-1}\mathbb{1}\{S_i=1\} \leqslant (p_{\min}-\eta)(t-1))$ is small. Let $\mathcal{F}_0 \coloneqq \{\varnothing, \Omega\}$ be the minimum σ-algebra, and $\mathcal{F}_i \coloneqq \sigma(S_1, A_1, ..., S_i, A_i)$ be the σ-algebra generated by the random variables up to time i. Since for any $a \in \mathcal{A}$,

$$\mathbb{P}\left(S_i = 1 | S_{i-1}=0, A_{i-1}=a\right) \geqslant \mu(a)(1-q(0,1))$$

$$\mathbb{P}\left(S_i = 1 | S_{i-1} = 1, A_{i-1} = a\right) \geqslant \mu(a)(1 - q(1,1)),$$

we have

$$\mathbb{E}\left[\mathbb{1}\{S_i = 1\} | \mathcal{F}_{i-1}\right] \geqslant \mu(a_M) \min\left\{1 - q(0,1), 1 - q(1,1)\right\} = p_{\min} > 0.$$

Hence we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{t-1} \mathbb{1}\{S_i = 1\} \leqslant (p_{\min} - \eta)(t-1)\right) = \mathbb{P}\left(\sum_{i=1}^{t-1} p_{\min} - \sum_{i=1}^{t-1} \mathbb{1}\{S_i = 1\} \geqslant \eta(t-1)\right) \\
\leqslant &\mathbb{P}\left(\sum_{i=1}^{t-1} \left(\mathbb{E}\left[\mathbb{1}\{S_i = 1\} | \mathcal{F}_{i-1}\right] - \mathbb{1}\{S_i = 1\}\right) \geqslant \eta(t-1)\right). \qquad (34)
\end{aligned}$$

Let $\Delta_i \coloneqq \mathbb{E}\left[\mathbb{1}\{S_i = 1\} | \mathcal{F}_{i-1}\right] - \mathbb{1}\{S_i = 1\}$. Note that Δ_i is measurable with respect to $\mathcal{F}_i$, $\mathbb{E}[\Delta_i | \mathcal{F}_{i-1}] = 0$, and $|\Delta_i| \leqslant 1$. Hence, by Azuma-Hoeffding inequality (Van Handel, 2016), we have

$$\mathbb{P}\left(\sum_{i=1}^{t-1} \left(\mathbb{E}\left[\mathbb{1}\{S_i = 1\} | \mathcal{F}_{i-1}\right] - \mathbb{1}\{S_i = 1\}\right) \geqslant \eta(t-1)\right) \leqslant \exp\left(-\frac{\eta^2(t-1)}{2}\right). \qquad (35)$$

Therefore, from (33), (34) and (35), it follows that

$$\begin{aligned}
&\mathbb{P}\left(N_t(a_1) \leqslant \frac{(p_{\min} - \eta)(t-1)}{2}\right) \\
\leqslant &\mathbb{P}\left(N_t^1(a_1) \leqslant \frac{(p_{\min} - \eta)(t-1)}{2}, \sum_{i=1}^{t-1} \mathbb{1}\{S_i = 1\} > (p_{\min} - \eta)(t-1)\right) + \exp\left(-\frac{\eta^2(t-1)}{2}\right) \\
\leqslant &\mathbb{P}\left(\sum_{i=2}^{M} N_t^1(a_i) > \frac{(p_{\min} - \eta)(t-1)}{2}\right) + \exp\left(-\frac{\eta^2(t-1)}{2}\right) \\
\leqslant &\mathbb{P}\left(N_t^1(a_j) > \frac{(p_{\min} - \eta)(t-1)}{2(M-1)}\right) + \exp\left(-\frac{\eta^2(t-1)}{2}\right), \qquad (36)
\end{aligned}$$

where the second inequality is due to the fact that $\sum_{i=1}^{M} N_t^1(a_i) = \sum_{i=1}^{t-1} \mathbb{1}\{S_i = 1\}$ and in the last inequality $j \in \operatorname{argmax}_{i \in \{2,\dots,M\}} N_t^1(a_i)$. Consider the event $\{N_t^1(a_j) > \frac{(p_{\min} - \eta)(t-1)}{2(M-1)}\}$. Let $\tau_t < t$ be the time step when a_j is pulled in state 1 for the $\left\lceil \frac{(p_{\min} - \eta)(t-1)}{2(M-1)} \right\rceil$-th time. Then we have

$$\tau_t \geqslant \left\lceil \frac{(p_{\min} - \eta)(t-1)}{2(M-1)} \right\rceil + (M-1) \geqslant \frac{(p_{\min} - \eta)(t-1)}{2(M-1)} + (M-1) \qquad (37)$$

$$N_{\tau_t}^1(a_j) = \left\lceil \frac{(p_{\min} - \eta)(t-1)}{2(M-1)} \right\rceil - 1 \geqslant \frac{(p_{\min} - \eta)(t-1)}{2(M-1)} - 1, \qquad (38)$$

where the first inequality is due to the fact that the ULCB algorithm pulls the other $(M-1)$ arms at the beginning. Let $L_t \coloneqq \frac{(p_{\min} - \eta)(t-1)}{2(M-1)} + (M-1)$. Then we have

$$\mathbb{P}\left(N_t^1(a_j) > \frac{(p_{\min} - \eta)(t-1)}{2(M-1)}\right)$$

$$
\begin{aligned}
\leqslant & \mathbb{P}\left(\tau_t \geqslant L_t, N^1_{\tau_t}(a_j) = \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1, S_{\tau_t} = 1, A_{\tau_t} = a_j\right) \\
\leqslant & \mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_j) \geqslant \tilde{\mu}^1_{\tau_t}(a_1), \tau_t \geqslant L_t, N^1_{\tau_t}(a_j) = \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
\leqslant & \mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_j) \geqslant \tilde{\mu}^1_{\tau_t}(a_1), \tau_t \geqslant L_t, N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
\leqslant & \mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_j) \geqslant \tilde{\mu}^1_{\tau_t}(a_1), \tilde{\mu}^1_{\tau_t}(a_1) \geqslant \mu(a_1), \tau_t \geqslant L_t, N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
& + \mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right), \qquad (39)
\end{aligned}
$$

where the first inequality follows from (37), (38), and the definition of τ_t, the second inequality follows from $S_{\tau_t} = 1$, $A_{\tau_t} = a_j$, and Line 13 of Algorithm 1, the third inequality is due to the fact that $N_{\tau_t}(a_j) \geqslant N^1_{\tau_t}(a_j)$, and the last inequality is by law of total probability.

By Lemma 12 (which is presented after this proof) and $c_1 = 1, c = 4$, for any $t \geqslant T_1$, we can bound the second term in (39) as follows

$$
\begin{aligned}
& \mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) \leqslant \frac{e\log L_t \log(t-2) + 4e\log(\log L_t)\log(t-2) + e}{L_t(\log L_t)^4} \\
\leqslant & \frac{e\log(t-2) + 4\log(t-2) + e}{L_t(\log L_t)^3} \leqslant \frac{(4+e)\log(t-1) + e}{\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}\left[\log\left(\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}\right)\right]^3} \\
\leqslant & \frac{c_3}{c_2(t-1)\left[\log\left(c_2(t-1)\right)\right]^2}, \qquad (40)
\end{aligned}
$$

where the second inequality is obtained by dividing the numerator and the denominator by $\log L_t$, the third inequality is by the definition of L_t, and the last inequality is obtained by dividing the numerator and the denominator by $\log\left(c_2(t-1)\right)$, where $c_2 := \frac{p_{\min}-\eta}{2(M-1)}$ and $c_3 := \frac{4+e}{\log 2}\log\frac{2}{c_2} + \frac{e}{\log 2}$. For the first term in (39), we have

$$
\begin{aligned}
& \mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_j) \geqslant \tilde{\mu}^1_{\tau_t}(a_1), \tilde{\mu}^1_{\tau_t}(a_1) \geqslant \mu(a_1), \tau_t \geqslant L_t, N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
\leqslant & \mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_j) \geqslant \mu(a_1), N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right). \qquad (41)
\end{aligned}
$$

Consider the event $\{\tilde{\mu}^1_{\tau_t}(a_j) \geqslant \mu(a_1), N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\}$. Then we have

$$
\begin{aligned}
\tilde{\mu}^1_{\tau_t}(a_j) = & \bar{\mu}_{\tau_t}(a_j) + \sqrt{\frac{\log\tau_t + 4\log\left(\log\tau_t\right)}{2N_{\tau_t}(a_j)}} \\
& \geqslant \mu(a_1) = \mu(a_j) + (\mu(a_1) - \mu(a_j)) \geqslant \mu(a_j) + (\mu(a_1) - \mu(a_2)),
\end{aligned}
$$

which implies

$$
\bar{\mu}_{\tau_t}(a_j) - \mu(a_j) \geqslant (\mu(a_1) - \mu(a_2)) - \sqrt{\frac{\log\tau_t + 4\log\left(\log\tau_t\right)}{2N_{\tau_t}(a_j)}}
$$

$$\geqslant(\mu(a_1)-\mu(a_2))-\sqrt{\frac{\log t+4\log(\log t)}{\frac{(p_{\min}-\eta)(t-1)}{(M-1)}-2}}\quad\geqslant\gamma,$$

where the second inequality is by $\tau_t < t$ and $N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1$, and the last inequality is by $t \geqslant T_1$ and the definition of T_1. Hence we have

$$\begin{aligned}
&\mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_j)\geqslant\mu(a_1), N_{\tau_t}(a_j)\geqslant\left\lceil\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}\right\rceil-1\right)\\
\leqslant&\mathbb{P}\left(\bar{\mu}_{\tau_t}(a_j)-\mu(a_j)\geqslant\gamma, N_{\tau_t}(a_j)\geqslant\left\lceil\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}\right\rceil-1\right)\\
\leqslant&\sum_{i=2}^{M}\mathbb{P}\left(\bar{\mu}_{\tau_t}(a_i)-\mu(a_i)\geqslant\gamma, N_{\tau_t}(a_i)\geqslant\left\lceil\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}\right\rceil-1\right)\\
\leqslant&\sum_{i=2}^{M}\sum_{n=\left\lceil\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}\right\rceil-1}^{t-1}\mathbb{P}\left(\frac{1}{n}\sum_{s=1}^{n}R_s(a_i)-\mu(a_i)\geqslant\gamma\right)\\
\leqslant&(M-1)\sum_{n=\left\lceil\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}\right\rceil-1}^{t-1}\exp(-2n\gamma^2)\leqslant\frac{M-1}{2\gamma^2\exp\left(2\gamma^2c_2(t-1)-4\gamma^2\right)}.
\end{aligned}\tag{42}$$

In the derivation above, the second inequality is by the union bound over all possible j. The third inequality is by the union bound over all possible number of pulls of arm a_i, where $\{R_s(a_i)\}_{s=1}^n$ are n i.i.d. Bernoulli rewards of pulling arm a_i. The fourth inequality uses Hoeffding inequality, and the last inequality is by integration.

From (36), (39), (40), (41), and (42), it follows that for any $t \geqslant T_1$,

$$\begin{aligned}
&\mathbb{P}\left(N_t(a_1)\leqslant\frac{(p_{\min}-\eta)(t-1)}{2}\right)\\
\leqslant&\frac{M-1}{2\gamma^2\exp\left(2\gamma^2c_2(t-1)-4\gamma^2\right)}+\frac{c_3}{c_2(t-1)\left[\log\left(c_2(t-1)\right)\right]^2}+\exp\left(-\frac{\eta^2(t-1)}{2}\right).
\end{aligned}$$

Lemma 12 *Let all the assumptions in Lemma 7 hold. Consider the ULCB algorithm. Let* $L_t \coloneqq \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} + (M-1)$, $\delta_{L_t} \coloneqq c_1^2 \log L_t + cc_1^2 \log(\log L_t)$. *Then for any* $t \geqslant T_1$,

$$\mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) \leqslant e\left\lceil\delta_{L_t}\log(t-2)\right\rceil\exp(-\delta_{L_t}) \tag{43}$$

Proof This lemma can be proved by a minor modification of the proof of Theorem 10 in the work of Garivier and Cappé (2011). The main difference is that τ_t is a random variable with a lower bound L_t. We present the whole proof for completeness. We first define several notations and construct a martingale. Let n be any time step. Let $\{X_i\}_{i=1}^n$ be the i.i.d. Bernoulli rewards generated by pulling arm a_1. Let $\{\mathcal{F}'_i\}$ be an increasing sequence of σ-algebra such that

$$\mathcal{F}'_0 \coloneqq \sigma(S_1, A_1), \quad \mathcal{F}'_i \coloneqq \sigma(S_1, A_1, X_1, ..., S_i, A_i, X_i, S_{i+1}, A_{i+1}),\ i \geqslant 1.$$

Note that X_i is independent of $\mathcal{F}'_{i-1}$ and X_i is measurable with respect to $\mathcal{F}'_i$. By the definition of τ_n, the event $\{\tau_n - 1 \geqslant i\} = \{\tau_n \leqslant i\}^{\mathrm{c}}$ is measurable with respect to $\mathcal{F}'_{i-1}$ for any $i \in \{1, ..., n-2\}$. Let

$$V_n \coloneqq \sum_{i=1}^{n-2} \epsilon_i X_i, \quad U_n \coloneqq \sum_{i=1}^{n-2} \epsilon_i, \quad n \geqslant 2,$$

where $\epsilon_i \coloneqq \mathbb{1}\{A_i = a_1, i \leqslant \tau_n - 1\}$, which is measurable with respect to $\mathcal{F}'_{i-1}$. Hence, V_n and U_{n+1} are measurable with respect to $\mathcal{F}'_{n-2}$. For any $\lambda \in \mathbb{R}$, let $\phi(\lambda) \coloneqq \log \mathbb{E}\left[\exp(\lambda X_1)\right]$. For any $n \geqslant 0$, define W_n^λ by

$$W_n^\lambda \coloneqq \exp\left(\lambda V_{n+2} - U_{n+2}\phi(\lambda)\right). \tag{44}$$

For any $n \geqslant 1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda(V_{n+2} - V_{n+1})\right)|\mathcal{F}'_{n-1}\right] = \mathbb{E}\left[\exp\left(\lambda\epsilon_n X_n\right)|\mathcal{F}'_{n-1}\right] = \mathbb{E}\left[\left(\exp\left(\lambda X_n\right)\right)^{\epsilon_n}|\mathcal{F}'_{n-1}\right] \\
&= \left(\mathbb{E}\left[\exp\left(\lambda X_n\right)|\mathcal{F}'_{n-1}\right]\right)^{\epsilon_n} = \exp\left(\epsilon_n \log \mathbb{E}\left[\exp(\lambda X_n)|\mathcal{F}'_{n-1}\right]\right) \\
&= \exp\left(\epsilon_n\phi(\lambda)\right) = \exp\left(\left(U_{n+2} - U_{n+1}\right)\phi(\lambda)\right),
\end{aligned} \tag{45}$$

where the third equality is due to the fact that $\epsilon_n \in \{0, 1\}$ and ϵ_n is measurable with respect to $\mathcal{F}'_{n-1}$, and the fifth equality is due to the fact that X_n is independent of $\mathcal{F}'_{n-1}$. Hence, by (45) and the fact that V_{n+1} and U_{n+2} are measurable with respect to $\mathcal{F}'_{n-1}$ we have

$$\mathbb{E}\left[\exp\left(\lambda V_{n+2} - U_{n+2}\phi(\lambda)\right)|\mathcal{F}'_{n-1}\right] = \exp\left(\lambda V_{n+1} - U_{n+1}\phi(\lambda)\right),$$

i.e., $\mathbb{E}\left[W_n^\lambda|\mathcal{F}'_{n-1}\right] = W_{n-1}^\lambda$, which implies that W_n^λ is a martingale with respect to the filtration $\{\mathcal{F}'_n\}$. Hence we have for any n,

$$\mathbb{E}\left[W_n^\lambda\right] = \mathbb{E}\left[W_0^\lambda\right] = 1. \tag{46}$$

Next we will use this conclusion to bound $\mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right)$. Note that $N_{\tau_t}(a_1) = U_t$ and $\bar{\mu}_{\tau_t}(a_1) = V_t/U_t$ by definition. By Line 12 of Algorithm 1, $\tilde{\mu}^1_{\tau_t}(a_1)$ can also be written as

$$\tilde{\mu}^1_{\tau_t}(a_1) = \max\left\{p : 2U_t\left(p - \bar{\mu}_{\tau_t}(a_1)\right)^2 \leqslant c_1^2 \log\tau_t + cc_1^2\log(\log\tau_t)\right\}. \tag{47}$$

Since $t \geqslant T_1$, we have $L_t \geqslant M + 1 \geqslant 3$ by the definition of L_t. Hence, by the definition of δ_{L_t}, we have $\delta_{L_t} > 0$. If $\delta_{L_t} \leqslant 1$, then

$$\mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) \leqslant 1 \leqslant \exp\left(1 - \delta_{L_t}\right)\lceil\delta_{L_t}\log(t-2)\rceil$$

for $t \geqslant T_1$. Then we only need to consider the case where $\delta_{L_t} > 1$. We use the same "peeling trick" as in the work of Garivier and Cappé (2011): we divide $\{1, 2, ..., t-2\}$ of possible values for U_t into slices $\{t_{n-1}+1, ..., t_{n-1}\}$ of geometrically increasing size, and treat the slices individually. Let $\beta_t \coloneqq \frac{1}{\delta_{L_t}-1}$. Let $t_0 \coloneqq 0$ and for $n \in \mathbb{N}$, $t_n \coloneqq \lfloor(1+\beta_t)^n\rfloor$. Let D_t be

the first integer such that $t_{D_t} \geqslant t-2$. Hence, $D_t = \left\lceil \frac{\log(t-2)}{\log(1+\beta_t)} \right\rceil$. Define $E_n := \{t_{n-1} < U_t \leqslant t_n\} \cap \{\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\}$. Then by union bound, we have

$$\mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) = \mathbb{P}\left(\bigcup_{n=1}^{D_t} E_n\right) \leqslant \sum_{n=1}^{D_t} \mathbb{P}\left(E_n\right). \tag{48}$$

Note that

$$\begin{aligned}
&\left\{\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right\} \\
\subseteq &\left\{\bar{\mu}_{\tau_t}(a_1) < \mu(a_1), 2U_t(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 > c_1^2 \log \tau_t + cc_1^2 \log(\log \tau_t), \tau_t \geqslant L_t\right\} \\
\subseteq &\left\{\bar{\mu}_{\tau_t}(a_1) < \mu(a_1), 2U_t(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 > \delta_{L_t}\right\}
\end{aligned} \tag{49}$$

by (47), $\tau_t \geqslant L_t$, and the definition of δ_{L_t}. Let m_t be the smallest integer such that $\frac{\delta_{L_t}}{m_t+1} \leqslant 2\mu(a_1)^2$. If $U_t \leqslant m_t$ and $\bar{\mu}_{\tau_t}(a_1) < \mu(a_1)$, then

$$2U_t(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 \leqslant 2m_t(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 \leqslant 2m_t\mu(a_1)^2 < \delta_{L_t},$$

which implies that

$$\left\{U_t \leqslant m_t, \bar{\mu}_{\tau_t}(a_1) < \mu(a_1), 2U_t(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 > \delta_{L_t}\right\} = \varnothing. \tag{50}$$

Therefore, it follows from (49) and (50) that

$$\left\{U_t \leqslant m_t, \tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right\} = \varnothing.$$

Hence, $E_n = \varnothing$ for all n such that $t_n \leqslant m_t$. For n such that $t_n > m_t$, let $\tilde{t}_{n-1} := \max\{t_{n-1}, m_t\}$. Then we have

$$\begin{aligned}
E_n \subseteq &\{\tilde{t}_{n-1} < U_t \leqslant t_n\} \cap \{\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\} \\
\subseteq &\{\tilde{t}_{n-1} < U_t \leqslant t_n\} \cap \{\bar{\mu}_{\tau_t}(a_1) < \mu(a_1), 2U_t(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 > \delta_{L_t}\},
\end{aligned} \tag{51}$$

where the second relation follows from (49). Define z_t such that $0 \leqslant z_t < \mu(a_1)$ and $2(\mu(a_1) - z_t)^2 = \frac{\delta_{L_t}}{(1+\beta_t)^n}$. Note that if E_n occurs, then the definition of z_t is valid since

$$\frac{\delta_{L_t}}{(1+\beta_t)^n} \leqslant \frac{\delta_{L_t}}{U_t} \leqslant \frac{\delta_{L_t}}{m_t+1} \leqslant 2\mu(a_1)^2,$$

where the first inequality follows from $U_t \leqslant t_n \leqslant (1+\beta_t)^n$, the second inequality follows from $U_t \geqslant m_t + 1$, and the third inequality is by the definition of m_t. For $U_t > \tilde{t}_{n-1}$, we have

$$\begin{aligned}
&E_n \cap \left\{U_t > \tilde{t}_{n-1}\right\} \subseteq E_n \cap \{U_t > t_{n-1}\} \subseteq E_n \cap \left\{U_t > \lfloor (1+\beta_t)^{n-1} \rfloor\right\} \\
\subseteq &E_n \cap \left\{U_t \geqslant (1+\beta_t)^{n-1}\right\} \subseteq E_n \cap \left\{2(\mu(a_1) - z_t)^2 = \frac{\delta_{L_t}}{(1+\beta_t)^n} \geqslant \frac{\delta_{L_t}}{(1+\beta_t)U_t}\right\},
\end{aligned} \tag{52}$$

where the first relation is by definition of $\tilde{t}_{n-1}$, the second relation is by the definition of t_{n-1}, and the last relation uses the definition of z_t. For $U_t \leqslant t_n$, we have

$$E_n \cap \{U_t \leqslant t_n\} \subseteq E_n \cap \{U_t \leqslant \lfloor (1+\beta_t)^n \rfloor\} \subseteq E_n \cap \{U_t \leqslant (1+\beta_t)^n\}$$

$$\begin{aligned}
&\subseteq E_n \cap \{U_t \leqslant (1+\beta_t)^n\} \cap \left\{2U_t(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 > \delta_{L_t}\right\} \\
&\subseteq E_n \cap \left\{2(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 > \frac{\delta_{L_t}}{U_t} \geqslant \frac{\delta_{L_t}}{(1+\beta_t)^n} = 2(\mu(a_1) - z_t)^2\right\}, \quad (53)
\end{aligned}$$

where the first relation is by the definition of t_n, the third relation follows from (51), and the last relation uses the definition of z_t. Hence, from (51) (52), and (53), it follows that

$$\begin{aligned}
E_n &\subseteq \left\{\bar{\mu}_{\tau_t}(a_1) < \mu(a_1), 2(\mu(a_1) - \bar{\mu}_{\tau_t}(a_1))^2 > 2(\mu(a_1) - z_t)^2 \geqslant \frac{\delta_{L_t}}{(1+\beta_t)U_t}\right\} \\
&\subseteq \left\{\bar{\mu}_{\tau_t}(a_1) < z_t, 2(\mu(a_1) - z_t)^2 \geqslant \frac{\delta_{L_t}}{(1+\beta_t)U_t}\right\}. \quad (54)
\end{aligned}$$

Define $\lambda_t \coloneqq \log\left(z_t(1-\mu(a_1))\right) - \log\left(\mu(a_1)(1-z_t)\right) \leqslant 0$. By Lemma 9 in the work of Garivier and Cappé (2011), we have

$$\phi(\lambda_t) \leqslant \log\left(1 - \mu(a_1) + \mu(a_1)\exp(\lambda_t)\right). \quad (55)$$

Then it follows from (54) and (55) that

$$\begin{aligned}
E_n \subseteq &\left\{\lambda_t\bar{\mu}_{\tau_t}(a_1) - \phi(\lambda_t) \geqslant \lambda_t z_t - \log\left(1 - \mu(a_1) + \mu(a_1)\exp(\lambda_t)\right),\right. \\
&\left. 2(\mu(a_1) - z_t)^2 \geqslant \frac{\delta_{L_t}}{(1+\beta_t)U_t}\right\} \\
= &\left\{\lambda_t\bar{\mu}_{\tau_t}(a_1) - \phi(\lambda_t) \geqslant \mathrm{kl}(z_t, \mu(a_1)), 2(\mu(a_1) - z_t)^2 \geqslant \frac{\delta_{L_t}}{(1+\beta_t)U_t}\right\} \\
\subseteq &\left\{\lambda_t\bar{\mu}_{\tau_t}(a_1) - \phi(\lambda_t) \geqslant 2(\mu(a_1) - z_t)^2, 2(\mu(a_1) - z_t)^2 \geqslant \frac{\delta_{L_t}}{(1+\beta_t)U_t}\right\} \\
\subseteq &\left\{\lambda_t\bar{\mu}_{\tau_t}(a_1) - \phi(\lambda_t) \geqslant \frac{\delta_{L_t}}{(1+\beta_t)U_t}\right\}, \quad (56)
\end{aligned}$$

where the second relation is by the definition of λ_t and $\mathrm{kl}(\cdot,\cdot)$, and the third relation uses Pinsker's inequality such that $2(\mu(a_1) - z_t)^2 \leqslant \mathrm{kl}(z_t, \mu(a_1))$. By the relation that $\bar{\mu}_{\tau_t}(a_1) = V_t/U_t$, the definition of W_n^λ in (44), and (56), we have

$$E_n \subseteq \left\{\log\left(W_{t-2}^{\lambda_t}\right) \geqslant \frac{\delta_{L_t}}{(1+\beta_t)}\right\} = \left\{W_{t-2}^{\lambda_t} \geqslant \exp\left(\frac{\delta_{L_t}}{(1+\beta_t)}\right)\right\}.$$

By Markov's inequality, we have

$$\mathbb{P}\left(E_n\right) \leqslant \mathbb{P}\left(W_{t-2}^{\lambda_t} \geqslant \exp\left(\frac{\delta_{L_t}}{(1+\beta_t)}\right)\right) \leqslant \frac{\mathbb{E}\left[W_{t-2}^{\lambda_t}\right]}{\exp\left(\frac{\delta_{L_t}}{(1+\beta_t)}\right)} = \exp\left(-\frac{\delta_{L_t}}{(1+\beta_t)}\right),$$

where the equality follows from (46). Hence, from (48), it follows that

$$\mathbb{P}\left(\tilde{\mu}_{\tau_t}^1(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) \leqslant \sum_{n=1}^{D_t} \mathbb{P}\left(E_n\right) \leqslant D_t \exp\left(-\frac{\delta_{L_t}}{(1+\beta_t)}\right). \quad (57)$$

Recall that $\beta_t = \frac{1}{\delta_{L_t}-1}$. Then

$$D_t = \left\lceil \frac{\log(t-2)}{\log(1+\beta_t)} \right\rceil = \left\lceil \frac{\log(t-2)}{\log(1+\frac{1}{\delta_{L_t}-1})} \right\rceil \leqslant \lceil \delta_{L_t} \log(t-2) \rceil , \tag{58}$$

where the last inequality follows from the fact that $\log(1+\frac{1}{x-1}) \geqslant \frac{1}{x}$ for any $x > 1$. From (57) and (58), it follows that

$$\begin{aligned}\mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) \leqslant & \lceil \delta_{L_t} \log(t-2) \rceil \exp\left(-\frac{\delta_{L_t}}{(1+\beta_t)}\right) \\ = & e \lceil \delta_{L_t} \log(t-2) \rceil \exp(-\delta_{L_t}).\end{aligned}$$

■

B.5 Proof of Lemma 8

By monotone convergence theorem and linearity of expectation, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 0, A_t = a_i\} \left[V^*(0) - Q^*(0,a_i)\right]\right] \\ =&\mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}\{t \leqslant T(K,\pi)\} \sum_{i=2}^{M} \mathbb{1}\{S_t = 0, A_t = a_i\} \left[V^*(0) - Q^*(0,a_i)\right]\right] \\ =&\sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{1}\{t \leqslant T(K,\pi)\} \sum_{i=2}^{M} \mathbb{1}\{S_t = 0, A_t = a_i\} \left[V^*(0) - Q^*(0,a_i)\right]\right] \\ \leqslant&\sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i=2}^{M} \mathbb{1}\{S_t = 0, A_t = a_i\} \left[V^*(0) - Q^*(0,a_i)\right]\right] \\ =&\sum_{t=1}^{\infty} \sum_{i=2}^{M} \left[V^*(0) - Q^*(0,a_i)\right] \mathbb{E}\left[\mathbb{1}\{S_t = 0, A_t = a_i\}\right] \\ =&\sum_{t=1}^{\infty} \sum_{i=2}^{M} \left[V^*(0) - Q^*(0,a_i)\right] \mathbb{P}\left(S_t = 0, A_t = a_i\right) \\ =&\sum_{i=2}^{M} \left[V^*(0) - Q^*(0,a_i)\right] \sum_{t=1}^{\infty} \mathbb{P}\left(S_t = 0, A_t = a_i\right), \end{aligned} \tag{59}$$

where $\mathbb{P}\left(S_t = 0, A_t = a_i\right)$ can be bounded by

$$\mathbb{P}\left(S_t = 0, A_t = a_i\right) \leqslant \mathbb{P}\left(S_t = 0, A_t = a_i, \mu(a_i) \geqslant \tilde{\mu}^0_t(a_i)\right) + \mathbb{P}\left(\mu(a_i) < \tilde{\mu}^0_t(a_i)\right).$$

Note that by definition of $\tilde{\mu}^0_t(a_i)$ we know

$$\tilde{\mu}^0_t(a_i) = \bar{\mu}_t(a_i) - \sqrt{\frac{\log t + 4\log(\log t)}{2N_t(a_i)}} = \min\left\{p : 2(\bar{\mu}_t(a_i) - p)^2 N_t(a_i) \leqslant \log t + 4\log(\log t)\right\}. \tag{60}$$

Hence, by Theorem 10 in the work of Garivier and Cappé (2011), we have

$$\mathbb{P}\left(\mu(a_i) < \tilde{\mu}_t^0(a_i)\right) \leqslant \frac{e\left[\left[\log t + 4\log(\log t)\right]\log t\right]}{t(\log t)^4} \leqslant \frac{6e}{t(\log t)^2} \tag{61}$$

for any $t \geqslant T_1$. Hence, we have

$$\mathbb{P}\left(S_t = 0, A_t = a_i\right) \leqslant \mathbb{P}\left(S_t = 0, A_t = a_i, \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right) + \frac{6e}{t(\log t)^2}$$

for any $t \geqslant T_1$. Similarly, we have

$$\begin{aligned}
&\mathbb{P}\left(S_t = 0, A_t = a_i\right) \\
\leqslant&\mathbb{P}\left(S_t = 0, A_t = a_i, \tilde{\mu}_t^1(a_1) \geqslant \mu(a_1), \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right) + \mathbb{P}\left(\tilde{\mu}_t^1(a_1) < \mu(a_1)\right) + \frac{6e}{t(\log t)^2} \\
\leqslant&\mathbb{P}\left(S_t = 0, A_t = a_i, \tilde{\mu}_t^1(a_1) \geqslant \mu(a_1), \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right) + \frac{12e}{t(\log t)^2}
\end{aligned} \tag{62}$$

for any $t \geqslant T_1$. Note that $\tilde{\mu}_t^1(a_1) = \tilde{\mu}_t^0(a_1) + 2\sqrt{\frac{\log t + 4\log(\log t)}{2N_t(a_1)}}$ by definition. Hence, we have

$$\begin{aligned}
&\mathbb{P}\left(S_t = 0, A_t = a_i, \tilde{\mu}_t^1(a_1) \geqslant \mu(a_1), \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right) \\
=&\mathbb{P}\left(S_t = 0, A_t = a_i, \tilde{\mu}_t^0(a_1) + 2\sqrt{\frac{\log t + 4\log(\log t)}{2N_t(a_1)}} \geqslant \mu(a_1), \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right) \\
=&\mathbb{P}\Bigg(S_t = 0, A_t = a_i, \tilde{\mu}_t^0(a_1) + 2\sqrt{\frac{\log t + 4\log(\log t)}{2N_t(a_1)}} \geqslant \mu(a_i) + (\mu(a_1) - \mu(a_i)), \\
&\qquad \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\Bigg) \\
\leqslant&\mathbb{P}\left(S_t = 0, A_t = a_i, \tilde{\mu}_t^0(a_1) + 2\sqrt{\frac{\log t + 4\log(\log t)}{2N_t(a_1)}} \geqslant \tilde{\mu}_t^0(a_i) + (\mu(a_1) - \mu(a_i))\right) \\
\leqslant&\mathbb{P}\left(S_t = 0, A_t = a_i, \tilde{\mu}_t^0(a_1) \geqslant \tilde{\mu}_t^0(a_i) + (\mu(a_1) - \mu(a_2)) - 2\sqrt{\frac{\log t + 4\log(\log t)}{2N_t(a_1)}}\right).
\end{aligned}$$

Then by Lemma 7, for any $t \geqslant T_1$, we have

$$\begin{aligned}
&\mathbb{P}\left(S_t = 0, A_t = a_i, \tilde{\mu}_t^1(a_1) \geqslant \mu(a_1), \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right) \\
\leqslant&\mathbb{P}\Bigg(S_t = 0, A_t = a_i, \tilde{\mu}_t^0(a_1) \geqslant \tilde{\mu}_t^0(a_i) + (\mu(a_1) - \mu(a_2)) - 2\sqrt{\frac{\log t + 4\log(\log t)}{2N_t(a_1)}}, \\
&\qquad N_t(a_1) > \frac{(p_{\min} - \eta)(t-1)}{2}\Bigg) + \mathbb{P}\left(N_t(a_1) \leqslant \frac{(p_{\min} - \eta)(t-1)}{2}\right) \\
\leqslant&\mathbb{P}\left(S_t = 0, A_t = a_i, \tilde{\mu}_t^0(a_1) > \tilde{\mu}_t^0(a_i) + (\mu(a_1) - \mu(a_2)) - 2\sqrt{\frac{\log t + 4\log(\log t)}{(p_{\min} - \eta)(t-1)}}\right)
\end{aligned}$$

$$+\frac{M-1}{2\gamma^2\exp\left(2\gamma^2c_2(t-1)-4\gamma^2\right)}+\frac{c_3}{c_2(t-1)\left(\log\left(c_2(t-1)\right)\right)^2}+\exp\left(-\frac{\eta^2(t-1)}{2}\right)$$

Define T_2 such that $T_2 \geqslant T_1$ and for any $t \geqslant T_2$, $\mu(a_1)-\mu(a_2) \geqslant 2\sqrt{\frac{\log t+4\log(\log t)}{(p_{\min}-\eta)(t-1)}}$. Then for any $t \geqslant T_2$, we have

$$\begin{aligned}
&\mathbb{P}\left(S_t=0, A_t=a_i, \tilde{\mu}_t^1(a_1) \geqslant \mu(a_1), \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right)\\
\leqslant&\mathbb{P}\left(S_t=0, A_t=a_i, \tilde{\mu}_t^0(a_1) > \tilde{\mu}_t^0(a_i)\right)\\
&+\frac{M-1}{2\gamma^2\exp\left(2\gamma^2c_2(t-1)-4\gamma^2\right)}+\frac{c_3}{c_2(t-1)\left(\log\left(c_2(t-1)\right)\right)^2}+\exp\left(-\frac{\eta^2(t-1)}{2}\right)\\
\leqslant&\frac{M-1}{2\gamma^2\exp\left(2\gamma^2c_2(t-1)-4\gamma^2\right)}+\frac{c_3}{c_2(t-1)\left(\log\left(c_2(t-1)\right)\right)^2}+\exp\left(-\frac{\eta^2(t-1)}{2}\right), \qquad (63)
\end{aligned}$$

where the last inequality follows from $\mathbb{P}\left(S_t=0, A_t=a_i, \tilde{\mu}_t^0(a_1) > \tilde{\mu}_t^0(a_i)\right)=0$ by the ULCB algorithm. Therefore, it follows from (62) and (63) that

$$\begin{aligned}
\sum_{t=T_2}^{\infty}\mathbb{P}\left(S_t=0, A_t=a_i\right) \leqslant& \sum_{t=T_2}^{\infty}\frac{M-1}{2\gamma^2\exp\left(2\gamma^2c_2(t-1)-4\gamma^2\right)}+\frac{c_3}{c_2(t-1)\left(\log\left(c_2(t-1)\right)\right)^2}\\
&+\exp\left(-\frac{\eta^2(t-1)}{2}\right)+\frac{12e}{t(\log t)^2}\\
\leqslant&\frac{(M-1)\exp(4\gamma^2-2\gamma^2c_2(T_2-2))}{4\gamma^4c_2}+\frac{c_3}{c_2\log[c_2(T_2-2)]}\\
&+\frac{2}{\eta^2}\exp\left(-\frac{\eta^2(T_2-2)}{2}\right)+\frac{12e}{\log(T_2-1)}. \qquad (64)
\end{aligned}$$

Hence, from (59), it follows that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{S_t=0, A_t=a_i\}\left[V^*(0)-Q^*(0,a_i)\right]\right]\\
\leqslant&\sum_{i=2}^{M}\left[V^*(0)-Q^*(0,a_i)\right]\left[T_2+\sum_{t=T_2}^{\infty}\mathbb{P}\left(S_t=0, A_t=a_i\right)\right] \leqslant c_4\sum_{i=2}^{M}\left[V^*(0)-Q^*(0,a_i)\right],
\end{aligned}$$

where the last inequality follows from (64) and

$$\begin{aligned}
c_4 \coloneqq& T_2+\frac{(M-1)\exp(4\gamma^2-2\gamma^2c_2(T_2-2))}{4\gamma^4c_2}+\frac{c_3}{c_2\log[c_2(T_2-2)]}+\frac{2}{\eta^2}\exp\left(-\frac{\eta^2(T_2-2)}{2}\right)\\
&+\frac{12e}{\log(T_2-1)}.
\end{aligned}$$

B.6 Proof of Lemma 9

By monotone convergence theorem and linearity of expectation, we have

$$\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{S_t=1, A_t=a_i\}\left[V^*(1)-Q^*(1,a_i)\right]\right]$$

$$
\begin{aligned}
=&\mathbb{E}\left[\sum_{t=1}^{\infty}\mathbb{1}\{t\leqslant T(K,\pi)\}\sum_{i=2}^{M}\mathbb{1}\{S_t=1,A_t=a_i\}\left[V^*(1)-Q^*(1,a_i)\right]\right]\\
=&\mathbb{E}\left[\sum_{i=2}^{M}\sum_{t=1}^{\infty}\mathbb{1}\{S_t=1,A_t=a_i,t\leqslant T(K,\pi)\}\left[V^*(1)-Q^*(1,a_i)\right]\right]\\
=&\sum_{i=2}^{M}\left[V^*(1)-Q^*(1,a_i)\right]\mathbb{E}\left[\sum_{t=1}^{\infty}\mathbb{1}\{S_t=1,A_t=a_i,t\leqslant T(K,\pi)\}\right]\\
\leqslant&\sum_{i=2}^{M}\left[V^*(1)-Q^*(1,a_i)\right]\left(M+\mathbb{E}\left[\sum_{t=M+1}^{\infty}\mathbb{1}\{S_t=1,A_t=a_i,t\leqslant T(K,\pi)\}\right]\right). \qquad (65)
\end{aligned}
$$

Let $\epsilon>0$. Let $B(t)\coloneqq\frac{1+\epsilon}{2(\mu(a_1)-\mu(a_i))^2}\left[\log t+4\log(\log t)\right]$. Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=M+1}^{\infty}\mathbb{1}\{S_t=1,A_t=a_i,t\leqslant T(K,\pi)\}\right]\\
=&\mathbb{E}\left[B(T(K,\pi))+\sum_{t=M+1}^{\infty}\mathbb{1}\{S_t=1,A_t=a_i,t\leqslant T(K,\pi),N_t^1(a_i)\geqslant B(T(K,\pi))\}\right]\\
\leqslant&\mathbb{E}\left[B(T(K,\pi))+\sum_{t=M+1}^{\infty}\mathbb{1}\{S_t=1,A_t=a_i,t\leqslant T(K,\pi),N_t(a_i)\geqslant B(T(K,\pi))\}\right]\\
\leqslant&\mathbb{E}\left[B(T(K,\pi))+\sum_{t=M+1}^{\infty}\mathbb{1}\{S_t=1,A_t=a_i,N_t(a_i)\geqslant B(t)\}\right]\\
=&\mathbb{E}[B(T(K,\pi))]+\sum_{t=M+1}^{\infty}\mathbb{P}\left(S_t=1,A_t=a_i,N_t(a_i)\geqslant B(t)\right), \qquad (66)
\end{aligned}
$$

where the first inequality is due to the fact that $N_t(a_i)\geqslant N_t^1(a_i)$, the second inequality is due to the fact that $B(t)$ is increasing in t, and the last equality is by monotone convergence theorem. For the second term, we have

$$
\begin{aligned}
&\sum_{t=M+1}^{\infty}\mathbb{P}\left(S_t=1,A_t=a_i,N_t(a_i)\geqslant B(t)\right)\\
\leqslant&\sum_{t=M+1}^{\infty}\mathbb{P}\left(\tilde{\mu}_t^1(a_1)<\mu(a_1)\right)+\sum_{t=M+1}^{\infty}\mathbb{P}\left(S_t=1,A_t=a_i,N_t(a_i)\geqslant B(t),\tilde{\mu}_t^1(a_1)\geqslant\mu(a_1)\right)\\
\leqslant&\sum_{t=M+1}^{\infty}\frac{6e}{t(\log t)^2}+\sum_{t=M+1}^{\infty}\mathbb{P}\left(S_t=1,A_t=a_i,N_t(a_i)\geqslant B(t),\tilde{\mu}_t^1(a_1)\geqslant\mu(a_1)\right)\\
\leqslant&\sum_{t=M+1}^{\infty}\frac{6e}{t(\log t)^2}+\sum_{t=M+1}^{\infty}\mathbb{P}\left(S_t=1,A_t=a_i,N_t(a_i)\geqslant B(t),\tilde{\mu}_t^1(a_i)\geqslant\tilde{\mu}_t^1(a_1)\geqslant\mu(a_1)\right)\\
\leqslant&\frac{6e}{\log M}+\sum_{t=M+1}^{\infty}\mathbb{P}\left(A_t=a_i,N_t(a_i)\geqslant B(t),\tilde{\mu}_t^1(a_i)\geqslant\mu(a_1)\right), \qquad (67)
\end{aligned}
$$

where the second inequality is by Theorem 10 in the work of Garivier and Cappé (2011), similar to (61) in the proof of Lemma 8. The third inequality holds since the ULCB algorithm pulls arm a_i in state 1 if and only if $\tilde{\mu}_t^1(a_i) \geqslant \tilde{\mu}_t^1(a_1)$ when $t \geqslant M+1$. Consider that the event $\{A_t = a_i, N_t(a_i) \geqslant B(t), \tilde{\mu}_t^1(a_i) \geqslant \mu(a_1)\}$ holds. Then we have

$$\mu(a_1) \leqslant \tilde{\mu}_t^1(a_i) = \bar{\mu}_t(a_i) + \sqrt{\frac{\log t + 4\log(\log t)}{2N_t(a_i)}} \leqslant \bar{\mu}_t(a_i) + \frac{\mu(a_1) - \mu(a_i)}{\sqrt{1+\epsilon}}, \tag{68}$$

where the last inequality is by $N_t(a_i) \geqslant B(t)$ and the definition of $B(t)$. Define $r_\epsilon(a_i) \in (\mu(a_i), \mu(a_1))$ such that

$$\mu(a_1) - r_\epsilon(a_i) = \frac{\mu(a_1) - \mu(a_i)}{\sqrt{1+\epsilon}}. \tag{69}$$

Then it follows from (68) and (69) that $\mu(a_1) \leqslant \bar{\mu}_t(a_i) + \mu(a_1) - r_\epsilon(a_i)$, which implies that $\bar{\mu}_t(a_i) \geqslant r_\epsilon(a_i)$. Hence, the second term in (67) can be bounded by

$$\begin{aligned}
&\sum_{t=M+1}^{\infty} \mathbb{P}\left(A_t = a_i, N_t(a_i) \geqslant B(t), \tilde{\mu}_t^1(a_i) \geqslant \mu(a_1)\right) \leqslant \sum_{t=M+1}^{\infty} \mathbb{P}\left(A_t = a_i, \bar{\mu}_t(a_i) \geqslant r_\epsilon(a_i)\right) \\
&= \sum_{t=M+1}^{\infty} \mathbb{P}\left(A_t = a_i, \bar{\mu}_t(a_i) - \mu(a_i) \geqslant r_\epsilon(a_i) - \mu(a_i)\right) \\
&= \sum_{t=M+1}^{\infty} \sum_{n=1}^{t-1} \mathbb{P}\left(A_t = a_i, N_t(a_i) = n, \frac{1}{n}\sum_{s=1}^{n} R_s(a_i) - \mu(a_i) \geqslant r_\epsilon(a_i) - \mu(a_i)\right) \\
&\leqslant \sum_{n=1}^{\infty} \sum_{t=n+1}^{\infty} \mathbb{P}\left(A_t = a_i, N_t(a_i) = n, \frac{1}{n}\sum_{s=1}^{n} R_s(a_i) - \mu(a_i) \geqslant r_\epsilon(a_i) - \mu(a_i)\right) \\
&\leqslant \sum_{n=1}^{\infty} \mathbb{P}\left(\frac{1}{n}\sum_{s=1}^{n} R_s(a_i) - \mu(a_i) \geqslant r_\epsilon(a_i) - \mu(a_i)\right),
\end{aligned} \tag{70}$$

where the second equality is by law of total probability, where $\{R_s(a_i)\}_{s=1}^n$ are i.i.d. Bernoulli rewards of pulling arm a_i, and the last inequality is due to the fact that $\{A_t = a_i, N_t(a_i) = n\}_{t=n+1}^{\infty}$ are mutually exclusive and the countable additivity of probability measure. Then by Hoeffding inequality and (70), we have

$$\begin{aligned}
&\sum_{t=M+1}^{\infty} \mathbb{P}\left(A_t = a_i, N_t(a_i) \geqslant B(t), \tilde{\mu}_t^1(a_i) \geqslant \mu(a_1)\right) \\
&\leqslant \sum_{n=1}^{\infty} \exp\left(-2n\left(r_\epsilon(a_i) - \mu(a_i)\right)^2\right) \leqslant \frac{1}{2\left(r_\epsilon(a_i) - \mu(a_i)\right)^2}.
\end{aligned} \tag{71}$$

Therefore, from (65), (66), (67), and (71), it follows that

$$\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 1, A_t = a_i\}\left[V^*(1) - Q^*(1, a_i)\right]\right]$$

$$\leqslant \sum_{i=2}^{M} \left[V^*(1) - Q^*(1,a_i)\right] \left[M + \mathbb{E}[B(T(K,\pi))] + \frac{6e}{\log M} + \frac{1}{2\left(r_\epsilon(a_i) - \mu(a_i)\right)^2}\right], \tag{72}$$

where

$$\begin{aligned}
\mathbb{E}[B(T(K,\pi))] =& \mathbb{E}\left[\frac{1+\epsilon}{2(\mu(a_1)-\mu(a_i))^2}\left[\log T(K,\pi) + 4\log(\log T(K,\pi))\right]\right] \\
\leqslant& \frac{1+\epsilon}{2(\mu(a_1)-\mu(a_i))^2}\left[\log \mathbb{E}\left[T(K,\pi)\right] + 4\log\left(\log \mathbb{E}\left[T(K,\pi)\right]\right)\right] \\
\leqslant& \frac{1+\epsilon}{2(\mu(a_1)-\mu(a_i))^2}\left[\log c_5 K + 4\log\left(\log(c_5 K)\right)\right],
\end{aligned} \tag{73}$$

where the first inequality is by Jensen's inequality, and the second inequality is by Lemma 13 (which is presented after this proof). Then it follows from (72) and (73) that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 1, A_t = a_i\}\left[V^*(1) - Q^*(1,a_i)\right]\right] \\
\leqslant& \sum_{i\neq 1} \frac{1+\epsilon}{2(\mu(a_1)-\mu(a_i))^2}\left(V^*(1) - Q^*(1,a_i)\right)\log K + o(\log K).
\end{aligned} \tag{74}$$

Lemma 13 *Let Assumption 1 hold. Then for any policy $\pi \in \Pi$, we have*

$$\mathbb{E}[T(K,\pi)] = \sum_{k=1}^{K} \mathbb{E}\left[I_k(\pi, S_{k,1}, \phi_{k,1})\right] \leqslant \sum_{k=1}^{K} \mathbb{E}\left[I_k(\pi^*, S_{k,1})\right] \leqslant c_5 K.$$

Proof Define a genie-aided (model-based) policy π'_* that maximizes the expected number of steps in one episode, i.e., for any $s \in \{0,1\}$, $\pi'_* := \operatorname{argmax}_\pi \mathbb{E}\left[I(\pi, s)\right]$, where π is taken over all policies and $I(\pi, s)$ is the number of steps in one episode given initial state s. We omit the subscript k in $I(\pi, s)$ since the distribution does not depend on the episode number k. Then we have

$$\mathbb{E}\left[I(\pi'_*,1)|A_{k,1} = a\right] = 1 + \mu_a(1-q(1,1))\mathbb{E}\left[I(\pi'_*,1)\right] + (1-\mu_a)(1-q(1,0))\mathbb{E}\left[I(\pi'_*,0)\right] \tag{75}$$

$$\mathbb{E}\left[I(\pi'_*,0)|A_{k,1} = a\right] = 1 + \mu_a(1-q(0,1))\mathbb{E}\left[I(\pi'_*,1)\right] + (1-\mu_a)(1-q(0,0))\mathbb{E}\left[I(\pi'_*,0)\right] \tag{76}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[I(\pi'_*,1)|A_{k,1} = a\right] - \mathbb{E}\left[I(\pi'_*,0)|A_{k,1} = a\right] \\
=& \mu_a(q(0,1)-q(1,1))\mathbb{E}\left[I(\pi'_*,1)\right] + (1-\mu_a)(q(0,0)-q(1,0))\mathbb{E}\left[I(\pi'_*,0)\right] \geqslant 0,
\end{aligned}$$

where the inequality follows from Assumption 1. Hence, we have

$$\begin{aligned}
&\mathbb{E}\left[I(\pi'_*,1)\right] - \mathbb{E}\left[I(\pi'_*,0)\right] = \max_a \mathbb{E}\left[I(\pi'_*,1)|A_{k,1} = a\right] - \max_a \mathbb{E}\left[I(\pi'_*,0)|A_{k,1} = a\right] \\
=& \max_a \mathbb{E}\left[I(\pi'_*,1)|A_{k,1} = a\right] - \mathbb{E}\left[I(\pi'_*,0)|A_{k,1} = a'\right]
\end{aligned}$$

$$\geqslant \mathbb{E}\left[I(\pi'_*,1)|A_{k,1}=a'\right]-\mathbb{E}\left[I(\pi'_*,0)|A_{k,1}=a'\right]\geqslant 0, \tag{77}$$

where $a' \coloneqq \operatorname{argmax}_a \mathbb{E}\left[I(\pi'_*,0)|A_{k,1}=a\right]$. Hence, from (75) and (76), we have

$$\begin{aligned}&\mathbb{E}\left[I(\pi'_*,1)|A_{k,1}=a_1\right]-\mathbb{E}\left[I(\pi'_*,1)|A_{k,1}=a_i\right]\\ =&(\mu_{a_1}-\mu_{a_i})\Big[(1-q(1,1))\mathbb{E}\left[I(\pi'_*,1)\right]-(1-q(1,0))\mathbb{E}\left[I(\pi'_*,0)\right]\Big]\geqslant 0,\end{aligned}$$

where the inequality follows from (77) and Assumption 1. Similarly,

$$\begin{aligned}&\mathbb{E}\left[I(\pi'_*,0)|A_{k,1}=a_1\right]-\mathbb{E}\left[I(\pi'_*,0)|A_{k,1}=a_i\right]\\ =&(\mu_{a_1}-\mu_{a_i})\Big[(1-q(0,1))\mathbb{E}\left[I(\pi'_*,1)\right]-(1-q(0,0))\mathbb{E}\left[I(\pi'_*,0)\right]\Big]\geqslant 0.\end{aligned}$$

Therefore, the policy π'_* is always pulling a_1, i.e., $\pi'_*=\pi^*$, which implies that

$$\mathbb{E}[T(K,\pi)]=\sum_{k=1}^{K}\mathbb{E}\left[I_k(\pi,S_{k,1},\phi_{k,1})\right]\leqslant\sum_{k=1}^{K}\mathbb{E}\left[I_k(\pi^*,S_{k,1})\right]\leqslant\sum_{k=1}^{K}\mathbb{E}\left[I(\pi^*,1)\right]\leqslant c_5K.$$

■

B.7 Proof of Theorem 4: Upper Bound for KL-ULCB

The proof idea is similar to that of Theorem 2. The proof is based on the regret decomposition in (14). We will first show that the number of optimal pulls scales linearly with t with high probability. Then, we will bound the expected regrets induced in state 0 and state 1 respectively. We first prove a lemma similar to Lemma 7 as follows.

Lemma 14 *Let all the assumptions in Theorem 4 hold. Consider the KL-ULCB algorithm with $c_0=c_1=1$, and $c=4$. Let $p_{\min}\coloneqq\mu(a_M)\min\{1-q(0,1),1-q(1,1)\}$. Let $\eta\in(0,p_{\min})$ be a constant. Let $\gamma'>0$ be a constant and $r_{\gamma'}\in(\mu(a_2),\mu(a_1))$ such that $\mathrm{kl}(r_{\gamma'},\mu(a_1))=\frac{\mathrm{kl}(\mu(a_2),\mu(a_1))}{1+\gamma'}$. Define T_3 such that for any $t\geqslant T_3$,*

$$\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}\geqslant 2,\ \text{and}\ \frac{\log t+4\log(\log t)}{\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}-1}\leqslant\frac{\mathrm{kl}(\mu(a_2),\mu(a_1))}{1+\gamma'}.$$

Then for any $t\geqslant T_3$

$$\begin{aligned}&\mathbb{P}\left(N_t(a_1)\leqslant\frac{(p_{\min}-\eta)(t-1)}{2}\right)\leqslant\mathbb{P}\left(N_t^1(a_1)\leqslant\frac{(p_{\min}-\eta)(t-1)}{2}\right)\leqslant\exp\left(-\frac{\eta^2(t-1)}{2}\right)\\ &+\frac{M-1}{\mathrm{kl}\left(r_{\gamma'},\mu(a_2)\right)\exp\left(\mathrm{kl}\left(r_{\gamma'},\mu(a_2)\right)\left[\frac{(p_{\min}-\eta)(t-1)}{2(M-1)}-2\right]\right)}+\frac{c_3}{c_2(t-1)\left[\log\left(c_2(t-1)\right)\right]^2},\end{aligned}$$

where $c_2\coloneqq\frac{p_{\min}-\eta}{2(M-1)}$ and $c_3\coloneqq\frac{4+e}{\log 2}\log\frac{2}{c_2}+\frac{e}{\log 2}$ are constants.

Proof Let $t \geqslant T_3$. Using the same argument as in the proof of Lemma 7, we have

$$\begin{aligned}
&\mathbb{P}\left(N_t(a_1) \leqslant \frac{(p_{\min}-\eta)(t-1)}{2}\right) \leqslant \mathbb{P}\left(N_t^1(a_1) \leqslant \frac{(p_{\min}-\eta)(t-1)}{2}\right) \\
\leqslant &\mathbb{P}\left(\tilde{\mu}_{\tau_t}^1(a_j) \geqslant \mu(a_1), N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
&+ \mathbb{P}\left(\tilde{\mu}_{\tau_t}^1(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) + \exp\left(-\frac{\eta^2(t-1)}{2}\right),
\end{aligned} \tag{78}$$

where $N_t^1(a)$ denotes the number of times arm $a \in \mathcal{A}$ was pulled in state 1 before time step t, $j \in \operatorname{argmax}_{i\in\{2,\dots,M\}} N_t^1(a_i)$, $\tau_t < t$ denotes the time step when a_j is pulled in state 1 for the $\left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil$-th time, and $L_t := \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} + (M-1)$. The second term in (78), $\mathbb{P}\left(\tilde{\mu}_{\tau_t}^1(a_1) < \mu(a_1), \tau_t \geqslant L_t\right)$, can be bounded by

$$\mathbb{P}\left(\tilde{\mu}_{\tau_t}^1(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) \leqslant e\left\lceil \delta_{L_t} \log(t-2)\right\rceil \exp(-\delta_{L_t}), \tag{79}$$

where $\delta_{L_t} := \log L_t + 4\log(\log L_t)$. The proof of (79) is omitted since it can be proved the same way as Lemma 12 by just replacing the Euclidean distance in the proof of Lemma 12 with KL divergence. Hence, we have

$$\mathbb{P}\left(\tilde{\mu}_{\tau_t}^1(a_1) < \mu(a_1), \tau_t \geqslant L_t\right) \leqslant \frac{c_3}{c_2(t-1)\left[\log\left(c_2(t-1)\right)\right]^2}, \tag{80}$$

where $c_2 := \frac{p_{\min}-\eta}{2(M-1)}$ and $c_3 := \frac{4+e}{\log 2}\log\frac{2}{c_2} + \frac{e}{\log 2}$. Consider that the event

$$\left\{\tilde{\mu}_{\tau_t}^1(a_j) \geqslant \mu(a_1), N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right\}$$

holds. Define $\mathrm{kl}^+(x,y) := \mathrm{kl}(x,y)\mathbb{1}\{x<y\}$. Then we have

$$\begin{aligned}
\mathrm{kl}^+(\bar{\mu}_{\tau_t}(a_j), \mu(a_1)) \leqslant &\mathrm{kl}^+(\bar{\mu}_{\tau_t}(a_j), \tilde{\mu}_{\tau_t}^1(a_j)) = \frac{\log\tau_t + 4\log(\log\tau_t)}{N_{\tau_t}(a_j)} \leqslant \frac{\log t + 4\log(\log t)}{\frac{(p_{\min}-\eta)(t-1)}{2(M-1)} - 1} \\
\leqslant &\frac{\mathrm{kl}(\mu(a_2), \mu(a_1))}{1+\gamma'},
\end{aligned}$$

where the last inequality is by $t \geqslant T_3$ and the definition of T_3. Hence, for the first term in (78), we have

$$\begin{aligned}
&\mathbb{P}\left(\tilde{\mu}_{\tau_t}^1(a_j) \geqslant \mu(a_1), N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
\leqslant &\mathbb{P}\left(\mathrm{kl}^+(\bar{\mu}_{\tau_t}(a_j), \mu(a_1)) \leqslant \frac{\mathrm{kl}(\mu(a_2), \mu(a_1))}{1+\gamma'}, N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
\leqslant &\sum_{i=2}^{M} \mathbb{P}\left(\mathrm{kl}^+(\bar{\mu}_{\tau_t}(a_i), \mu(a_1)) \leqslant \frac{\mathrm{kl}(\mu(a_2), \mu(a_1))}{1+\gamma'}, N_{\tau_t}(a_i) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
\leqslant &\sum_{i=2}^{M} \sum_{n=\left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1}^{t-1} \mathbb{P}\left(\mathrm{kl}^+\left(\frac{1}{n}\sum_{s=1}^{n} R_s(a_i), \mu(a_1)\right) \leqslant \frac{\mathrm{kl}(\mu(a_2), \mu(a_1))}{1+\gamma'}\right),
\end{aligned} \tag{81}$$

where the second inequality is by the union bound over all possible j, and the last inequality is by the union bound over all possible number of pulls of arm a_i, where $\{R_s(a_i)\}_{s=1}^n$ are n i.i.d. Bernoulli rewards of pulling arm a_i. Consider that the event $\{\mathrm{kl}^+(\frac{1}{n}\sum_{s=1}^n R_s(a_i), \mu(a_1)) \leqslant \frac{\mathrm{kl}(\mu(a_2),\mu(a_1))}{1+\gamma'}\}$ holds. By the definition of $r_{\gamma'}$, we have $\mathrm{kl}^+(\frac{1}{n}\sum_{s=1}^n R_s(a_i), \mu(a_1)) \leqslant \mathrm{kl}(r_{\gamma'}, \mu(a_1))$, which implies that $\frac{1}{n}\sum_{s=1}^n R_s(a_i) \geqslant r_{\gamma'}$. Hence, we have

$$\mathrm{kl}\left(\frac{1}{n}\sum_{s=1}^{n} R_s(a_i), \mu(a_i)\right) \geqslant \mathrm{kl}\left(r_{\gamma'}, \mu(a_i)\right) \geqslant \mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right) \tag{82}$$

since $\mu(a_i) \leqslant \mu(a_2) < r_{\gamma'} \leqslant \frac{1}{n}\sum_{s=1}^n R_s(a_i)$. Hence, from (81) and (82), it follows that

$$\begin{aligned}
&\mathbb{P}\left(\tilde{\mu}^1_{\tau_t}(a_j) \geqslant \mu(a_1), N_{\tau_t}(a_j) \geqslant \left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1\right) \\
\leqslant &\sum_{i=2}^{M} \sum_{n=\left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1}^{t-1} \mathbb{P}\left(\mathrm{kl}\left(\frac{1}{n}\sum_{s=1}^{n} R_s(a_i), \mu(a_i)\right) \geqslant \mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right), \frac{1}{n}\sum_{s=1}^{n} R_s(a_i) > \mu(a_i)\right) \\
\leqslant &(M-1) \sum_{n=\left\lceil \frac{(p_{\min}-\eta)(t-1)}{2(M-1)} \right\rceil - 1}^{t-1} \exp\left(-n\mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right)\right) \\
\leqslant &\frac{M-1}{\mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right) \exp\left(\mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right)\left[\frac{(p_{\min}-\eta)(t-1)}{2(M-1)} - 2\right]\right)},
\end{aligned} \tag{83}$$

where the second inequality uses the concentration inequality for KL divergence (Mardia et al., 2020), and the last inequality is by integration. From (78), (80), and (83), Lemma 14 is proved. ■

Lemma 14 shows that when t is large enough, the number of optimal pulls scales linearly with t with high probability. This plays an important role in the proof of Lemma 15, which bounds the expected regret induced by pulling suboptimal arms in state 0 by a constant.

Lemma 15 *Let all the assumptions in Theorem 4 hold. Consider the KL-ULCB algorithm with $c_0 = c_1 = 1$, and $c = 4$. The regret induced in state 0 can be bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 0, A_t = a_i\}\left[V^*(0) - Q^*(0, a_i)\right]\right] \leqslant \sum_{i=2}^{M} c_{6,i}\left[V^*(0) - Q^*(0, a_i)\right],$$

where $c_{6,i}$ are constants which depend only on M, $\mu(a_1)$, $\mu(a_2)$, $\mu(a_i)$, $p_{\min}$, η, and γ'.

Proof Using the same argument as in the proof of Lemma 8, we have

$$\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 0, A_t = a_i\}\left[V^*(0) - Q^*(0, a_i)\right]\right]$$

$$\leqslant \sum_{i=2}^{M} \left[V^*(0) - Q^*(0, a_i)\right] \sum_{t=1}^{\infty} \mathbb{P}\left(S_t = 0, A_t = a_i\right), \tag{84}$$

where $\mathbb{P}\left(S_t = 0, A_t = a_i\right)$ can be bounded by

$$\begin{aligned}
&\mathbb{P}\left(S_t = 0, A_t = a_i\right) \leqslant \mathbb{P}\left(S_t = 0, A_t = a_i, \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right) + \mathbb{P}\left(\mu(a_i) < \tilde{\mu}_t^0(a_i)\right) \\
\leqslant &\mathbb{P}\left(S_t = 0, A_t = a_i, \mu(a_i) \geqslant \tilde{\mu}_t^0(a_i)\right) + \frac{6e}{t\left(\log t\right)^2}
\end{aligned}$$

for any $t \geqslant T_3$, where the inequality is by Theorem 10 in the work of Garivier and Cappé (2011). Note that $S_t = 0, A_t = a_i$ implies that $\tilde{\mu}_t^0(a_i) \geqslant \tilde{\mu}_t^0(a_1)$ by the KL-ULCB algorithm. Hence, for any $t \geqslant T_3$, we have

$$\begin{aligned}
&\mathbb{P}\left(S_t = 0, A_t = a_i\right) \leqslant \mathbb{P}\left(\mu(a_i) \geqslant \tilde{\mu}_t^0(a_i), \tilde{\mu}_t^0(a_i) \geqslant \tilde{\mu}_t^0(a_1)\right) + \frac{6e}{t\left(\log t\right)^2} \\
\leqslant &\mathbb{P}\left(\mu(a_i) \geqslant \tilde{\mu}_t^0(a_1)\right) + \frac{6e}{t\left(\log t\right)^2} \\
\leqslant &\mathbb{P}\left(\mu(a_i) \geqslant \tilde{\mu}_t^0(a_1), N_t(a_1) > \frac{(p_{\min} - \eta)(t-1)}{2}\right) + \mathbb{P}\left(N_t(a_1) \leqslant \frac{(p_{\min} - \eta)(t-1)}{2}\right) \\
&+ \frac{6e}{t\left(\log t\right)^2} \\
\leqslant &\mathbb{P}\left(\mu(a_i) \geqslant \tilde{\mu}_t^0(a_1), N_t(a_1) > \frac{(p_{\min} - \eta)(t-1)}{2}\right) + \frac{6e}{t\left(\log t\right)^2} + \exp\left(-\frac{\eta^2(t-1)}{2}\right) \\
&+ \frac{M-1}{\mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right) \exp\left(\mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right)\left[\frac{(p_{\min}-\eta)(t-1)}{2(M-1)} - 2\right]\right)} + \frac{c_3}{c_2(t-1)\left[\log\left(c_2(t-1)\right)\right]^2},
\end{aligned} \tag{85}$$

where the last inequality uses Lemma 14. Consider that the event $\{\mu(a_i) \geqslant \tilde{\mu}_t^0(a_1), N_t(a_1) > \frac{(p_{\min}-\eta)(t-1)}{2}\}$ holds. Define $\mathrm{kl}^-(x, y) \coloneqq \mathrm{kl}(x, y)\mathbb{1}\{x > y\}$. Hence, we have

$$\mathrm{kl}^-(\bar{\mu}_t(a_1), \mu(a_i)) \leqslant \mathrm{kl}(\bar{\mu}_t(a_1), \tilde{\mu}_t^0(a_1)) \leqslant \frac{\log t + 4\log(\log t)}{N_t(a_1)} \leqslant \frac{\log t + 4\log(\log t)}{\frac{(p_{\min}-\eta)(t-1)}{2}}.$$

Define T_4 such that $T_4 \geqslant T_3$ and for any $t \geqslant T_4$, $\frac{\log t + 4\log(\log t)}{\frac{(p_{\min}-\eta)(t-1)}{2}} \leqslant \frac{\mathrm{kl}(\mu(a_1), \mu(a_2))}{1+\gamma'}$. Then for any $t \geqslant T_4$, we have

$$\mathrm{kl}^-(\bar{\mu}_t(a_1), \mu(a_i)) \leqslant \frac{\mathrm{kl}(\mu(a_1), \mu(a_2))}{1+\gamma'} \leqslant \frac{\mathrm{kl}(\mu(a_1), \mu(a_i))}{1+\gamma'}.$$

Define $r'_{\gamma'}(a_i) \in (\mu(a_i), \mu(a_1))$ such that $\mathrm{kl}(r'_{\gamma'}(a_i), \mu(a_i)) = \frac{\mathrm{kl}(\mu(a_1), \mu(a_i))}{1+\gamma'}$. Then we have

$$\mathrm{kl}^-(\bar{\mu}_t(a_1), \mu(a_i)) \leqslant \mathrm{kl}(r'_{\gamma'}(a_i), \mu(a_i)),$$

which implies that $\bar{\mu}_t(a_1) \leqslant r'_{\gamma'}(a_i)$. Therefore, we have

$$\mathrm{kl}(\bar{\mu}_t(a_1), \mu(a_1)) \geqslant \mathrm{kl}(r'_{\gamma'}(a_i), \mu(a_1))$$

since $\bar{\mu}_t(a_1) \leqslant r'_{\gamma'}(a_i) < \mu(a_1)$. Hence, the first term in (85) can be bounded by

$$
\begin{aligned}
&\mathbb{P}\left(\mu(a_i) \geqslant \tilde{\mu}_t^0(a_1), N_t(a_1) > \frac{(p_{\min}-\eta)(t-1)}{2}\right) \\
\leqslant &\mathbb{P}\left(\mathrm{kl}(\bar{\mu}_t(a_1), \mu(a_1)) \geqslant \mathrm{kl}(r'_{\gamma'}(a_i), \mu(a_1)), N_t(a_1) > \frac{(p_{\min}-\eta)(t-1)}{2}, \bar{\mu}_t(a_1) < \mu(a_1)\right) \\
\leqslant &\sum_{n=\left\lceil \frac{(p_{\min}-\eta)(t-1)}{2} \right\rceil}^{t-1} \mathbb{P}\left(\mathrm{kl}\left(\frac{1}{n}\sum_{s=1}^{n} R_s(a_1), \mu(a_1)\right) \geqslant \mathrm{kl}\left(r'_{\gamma'}(a_i), \mu(a_1)\right), \frac{1}{n}\sum_{s=1}^{n} R_s(a_1) < \mu(a_1)\right) \\
\leqslant &\sum_{n=\left\lceil \frac{(p_{\min}-\eta)(t-1)}{2} \right\rceil}^{t-1} \exp\left(-n\mathrm{kl}\left(r'_{\gamma'}(a_i), \mu(a_1)\right)\right) \\
\leqslant &\frac{1}{\mathrm{kl}\left(r'_{\gamma'}(a_i), \mu(a_1)\right) \exp\left(\mathrm{kl}\left(r'_{\gamma'}(a_i), \mu(a_1)\right)\left[\frac{(p_{\min}-\eta)(t-1)}{2} - 1\right]\right)},
\end{aligned} \tag{86}
$$

where the second inequality is by the union bound over all possible number of pulls of arm a_1, where $\{R_s(a_1)\}_{s=1}^n$ are n i.i.d. Bernoulli rewards of pulling arm a_1, the third inequality uses the concentration inequality for KL divergence (Mardia et al., 2020), and the last inequality is by integration. Then by combining (85) and (86), we have

$$
\begin{aligned}
&\sum_{t=T_4}^{\infty} \mathbb{P}\left(S_t = 0, A_t = a_i\right) \\
\leqslant &\sum_{t=T_4}^{\infty} \frac{1}{\mathrm{kl}\left(r'_{\gamma'}(a_i), \mu(a_1)\right) \exp\left(\mathrm{kl}\left(r'_{\gamma'}(a_i), \mu(a_1)\right)\left[\frac{(p_{\min}-\eta)(t-1)}{2} - 1\right]\right)} + \frac{6e}{t\left(\log t\right)^2} \\
&\quad + \exp\left(-\frac{\eta^2(t-1)}{2}\right) + \frac{M-1}{\mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right) \exp\left(\mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right)\left[\frac{(p_{\min}-\eta)(t-1)}{2(M-1)} - 2\right]\right)} \\
&\quad + \frac{c_3}{c_2(t-1)\left[\log\left(c_2(t-1)\right)\right]^2} \\
\leqslant &\frac{2}{(p_{\min}-\eta)\left[\mathrm{kl}\left(r'_{\gamma'}(a_i), \mu(a_1)\right)\right]^2} + \frac{6e}{\log(T_4-1)} + \frac{2}{\eta^2} + \frac{2(M-1)^2}{\left[\mathrm{kl}\left(r_{\gamma'}, \mu(a_2)\right)\right]^2 (p_{\min}-\eta)} \\
&\quad + \frac{c_3}{c_2 \log[c_2(T_4-2)]}.
\end{aligned} \tag{87}
$$

Combining (84) and (87), we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 0, A_t = a_i\}\left[V^*(0) - Q^*(0, a_i)\right]\right] \\
\leqslant &\sum_{i=2}^{M} \left[V^*(0) - Q^*(0, a_i)\right]\left[T_4 + \sum_{t=T_4}^{\infty} \mathbb{P}\left(S_t = 0, A_t = a_i\right)\right] \leqslant \sum_{i=2}^{M} \left[V^*(0) - Q^*(0, a_i)\right] c_{6,i},
\end{aligned}
$$

where $c_{6,i} := \frac{2}{(p_{\min}-\eta)\left[\mathrm{kl}\left(r'_{\gamma'}(a_i),\mu(a_1)\right)\right]^2} + \frac{6e}{\log(T_4-1)} + \frac{2}{\eta^2} + \frac{2(M-1)^2}{\left[\mathrm{kl}\left(r_{\gamma'},\mu(a_2)\right)\right]^2(p_{\min}-\eta)} + \frac{c_3}{c_2\log[c_2(T_4-2)]}$. ■

Next, we will bound the expected regret induced by pulling suboptimal arms in state 1. Lemma 16 shows the result.

Lemma 16 *Let all the assumptions in Theorem 4 hold. Consider the KL-ULCB algorithm with $c_0 = c_1 = 1$, and $c = 4$. For any $\epsilon > 0$, the regret induced in state 1 can be bounded by*

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{S_t=1,A_t=a_i\}\left[V^*(1)-Q^*(1,a_i)\right]\right]\\ \leqslant&\sum_{i\neq 1}\frac{1+\epsilon}{\mathrm{kl}(\mu(a_i),\mu(a_1))}\left(V^*(1)-Q^*(1,a_i)\right)\log K+o(\log K)\end{aligned}\tag{88}$$

Proof This proof is similar to the proof of Lemma 9. Using the same argument as in the proof of Lemma 9, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)}\sum_{i=2}^{M}\mathbb{1}\{S_t=1,A_t=a_i\}\left[V^*(1)-Q^*(1,a_i)\right]\right]\\ \leqslant&\sum_{i=2}^{M}\left[V^*(1)-Q^*(1,a_i)\right]\\ &\left(M+\mathbb{E}[B'(T(K,\pi))]+\frac{6e}{\log M}+\sum_{t=M+1}^{\infty}\mathbb{P}\left(A_t=a_i,N_t(a_i)\geqslant B'(t),\tilde{\mu}_t^1(a_i)\geqslant\mu(a_1)\right)\right),\end{aligned}\tag{89}$$

where $B'(t) := \frac{1+\epsilon}{\mathrm{kl}(\mu(a_i),\mu(a_1))}\left[\log t + 4\log(\log t)\right]$. Consider that the event $\{A_t = a_i, N_t(a_i) \geqslant B'(t), \tilde{\mu}_t^1(a_i) \geqslant \mu(a_1)\}$ holds. Then we have

$$\begin{aligned}\mathrm{kl}^+\left(\bar{\mu}_t(a_i),\mu(a_1)\right)\leqslant&\mathrm{kl}\left(\bar{\mu}_t(a_i),\tilde{\mu}_t^1(a_i)\right)\leqslant\frac{\log t+4\log(\log t)}{N_t(a_i)}\\ \leqslant&\frac{\log t+4\log(\log t)}{B'(t)}=\frac{\mathrm{kl}(\mu(a_i),\mu(a_1))}{1+\epsilon},\end{aligned}$$

where $\mathrm{kl}^+(x,y) := \mathrm{kl}(x,y)\mathbb{1}\{x<y\}$. Define $r'_\epsilon(a_i) \in \{\mu(a_i),\mu(a_1)\}$ such that $\mathrm{kl}(r'_\epsilon(a_i),\mu(a_1)) = \frac{\mathrm{kl}(\mu(a_i),\mu(a_1))}{1+\epsilon}$. Then we have

$$\mathrm{kl}^+\left(\bar{\mu}_t(a_i),\mu(a_1)\right)\leqslant\mathrm{kl}(r'_\epsilon(a_i),\mu(a_1)),$$

which implies that $\mu(a_i) \leqslant r'_\epsilon(a_i) \leqslant \bar{\mu}_t(a_i)$. Hence we have

$$\mathrm{kl}(\bar{\mu}_t(a_i),\mu(a_i))\geqslant\mathrm{kl}(r'_\epsilon(a_i),\mu(a_i)).$$

Therefore, we have

$$\begin{aligned}
&\sum_{t=M+1}^{\infty} \mathbb{P}\left(A_t = a_i, N_t(a_i) \geqslant B'(t), \tilde{\mu}_t^1(a_i) \geqslant \mu(a_1)\right) \\
\leqslant &\sum_{t=M+1}^{\infty} \mathbb{P}\left(A_t = a_i, \mathrm{kl}(\bar{\mu}_t(a_i), \mu(a_i)) \geqslant \mathrm{kl}(r'_\epsilon(a_i), \mu(a_i))\right) \\
= &\sum_{t=M+1}^{\infty} \sum_{n=1}^{t-1} \mathbb{P}\left(A_t = a_i, N_t(a_i) = n, \mathrm{kl}\left(\frac{1}{n}\sum_{s=1}^{n} R_s(a_i), \mu(a_i)\right) \geqslant \mathrm{kl}\left(r'_\epsilon(a_i), \mu(a_i)\right)\right) \\
\leqslant &\sum_{n=1}^{\infty} \sum_{t=n+1}^{\infty} \mathbb{P}\left(A_t = a_i, N_t(a_i) = n, \mathrm{kl}\left(\frac{1}{n}\sum_{s=1}^{n} R_s(a_i), \mu(a_i)\right) \geqslant \mathrm{kl}\left(r'_\epsilon(a_i), \mu(a_i)\right)\right) \\
\leqslant &\sum_{n=1}^{\infty} \mathbb{P}\left(\mathrm{kl}\left(\frac{1}{n}\sum_{s=1}^{n} R_s(a_i), \mu(a_i)\right) \geqslant \mathrm{kl}\left(r'_\epsilon(a_i), \mu(a_i)\right)\right) \\
\leqslant &\sum_{n=1}^{\infty} \exp\left(-n\mathrm{kl}\left(r'_\epsilon(a_i), \mu(a_i)\right)\right) \leqslant \frac{1}{\mathrm{kl}\left(r'_\epsilon(a_i), \mu(a_i)\right)},
\end{aligned} \tag{90}$$

where the first equality is by law of total probability, where $\{R_s(a_i)\}_{s=1}^n$ are i.i.d. Bernoulli rewards of pulling arm a_i. The third inequality is due to the fact that $\{A_t = a_i, N_t(a_i) = n\}_{t=n+1}^{\infty}$ are mutually exclusive and the countable additivity of probability measure. The fourth inequality uses the concentration inequality for KL divergence (Mardia et al., 2020), and the last inequality is by integration. It then follows from (89) and (90) that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 1, A_t = a_i\}\left[V^*(1) - Q^*(1, a_i)\right]\right] \\
\leqslant &\sum_{i=2}^{M} \left[V^*(1) - Q^*(1, a_i)\right]\left(M + \mathbb{E}[B'(T(K,\pi))] + \frac{6e}{\log M} + \frac{1}{\mathrm{kl}\left(r'_\epsilon(a_i), \mu(a_i)\right)}\right),
\end{aligned} \tag{91}$$

where

$$\begin{aligned}
\mathbb{E}[B'(T(K,\pi))] =&\mathbb{E}\left[\frac{1+\epsilon}{\mathrm{kl}(\mu(a_i), \mu(a_1))}\left[\log T(K,\pi) + 4\log(\log T(K,\pi))\right]\right] \\
\leqslant&\frac{1+\epsilon}{\mathrm{kl}(\mu(a_i), \mu(a_1))}\left[\log \mathbb{E}\left[T(K,\pi)\right] + 4\log\left(\log \mathbb{E}\left[T(K,\pi)\right]\right)\right] \\
\leqslant&\frac{1+\epsilon}{\mathrm{kl}(\mu(a_i), \mu(a_1))}\left[\log c_5 K + 4\log\left(\log(c_5 K)\right)\right],
\end{aligned} \tag{92}$$

where the first inequality is by Jensen's inequality, and the second inequality is by Lemma 13. It then follows from (91) and (92) that

$$\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 1, A_t = a_i\}\left[V^*(1) - Q^*(1, a_i)\right]\right]$$

$$\leqslant \sum_{i\neq 1} \frac{1+\epsilon}{\mathrm{kl}(\mu(a_i),\mu(a_1))} \left(V^*(1) - Q^*(1,a_i)\right) \log K + o(\log K).$$

■

By the regret decomposition result (14), Lemma 15, and Lemma 16, we have

$$\limsup_{K\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_\pi(K)]}{\log K} \leqslant \sum_{i\neq 1} \frac{1+\epsilon}{\mathrm{kl}(\mu(a_i),\mu(a_1))} \left(V^*(1) - Q^*(1,a_i)\right),$$

i.e., Theorem 4 is proved.

B.8 Proof of Theorem 6: Lower Bound

From the regret decomposition (14), given any consistent policy $\pi \in \Pi_{\mathrm{cons}}$, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_\pi(K)] = \mathbb{E}\Bigg[&\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{S_t = 0, A_t = a_i\} \left(V^*(0) - Q^*(0,a_i)\right) \\
&+ \mathbb{1}\{S_t = 1, A_t = a_i\} \left(V^*(1) - Q^*(1,a_i)\right)\Bigg] \\
\geqslant \mathbb{E}&\left[\sum_{t=1}^{T(K,\pi)} \sum_{i=2}^{M} \mathbb{1}\{A_t = a_i\} \left(V^*(1) - Q^*(1,a_i)\right) \right] \\
= \sum_{i=2}^{M} &\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \mathbb{1}\{A_t = a_i\} \right] \left(V^*(1) - Q^*(1,a_i)\right),
\end{aligned} \tag{93}$$

where the first inequality uses the conclusion of Lemma 3, $V^*(0) - Q^*(0,a_i) \geqslant V^*(1) - Q^*(1,a_i)$. We have to bound the term $\mathbb{E}\left[\sum_{t=1}^{T(K,\pi)} \mathbb{1}\{A_t = a_i\}\right] = \mathbb{E}\left[N_{T(K,\pi)+1}(a_i)\right]$. Let $\epsilon \in (0,1)$. Consider a different system i, $i \in \{2,...,M\}$, where the only difference is that the mean value of the reward of pulling arm a_i is $\mu'(a_i) \in (\mu(a_1),1)$ such that

$$\mathrm{kl}(\mu(a_i),\mu'(a_i)) \leqslant (1+\epsilon)\mathrm{kl}(\mu(a_i),\mu(a_1)). \tag{94}$$

Define event C_i as follows

$$C_i := \left\{ N_{T(K,\pi)+1}(a_i) \leqslant \frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))} \log K,\ \hat{\mathrm{kl}}_{N_{T(K,\pi)+1}(a_i)} \leqslant (1-\frac{\epsilon}{2}) \log K \right\},$$

where for any n,

$$\hat{\mathrm{kl}}_n := \sum_{s=1}^{n} \log \frac{\mu(a_i)R_s(a_i) + (1-\mu(a_i))(1-R_s(a_i))}{\mu'(a_i)R_s(a_i) + (1-\mu'(a_i))(1-R_s(a_i))}, \tag{95}$$

where $\{R_s(a_i)\}_{s=1}^n$ are i.i.d. Bernoulli rewards of pulling arm a_i in the original system. It can be easily verified that $\mathbb{E}[\hat{\mathrm{kl}}_n] = n\mathrm{kl}(\mu(a_i),\mu'(a_i))$. We first show that the change of measure identity (96) holds.

Lemma 17 *Given a policy $\pi \in \Pi$ and total number of episodes K, we have*

$$\mathbb{P}'(C_i) = \mathbb{E}\left[\mathbb{1}_{C_i} \exp\left(-\hat{\mathrm{kl}}_{N_{T(K,\pi)+1}(a_i)}\right)\right], \tag{96}$$

where $\mathbb{P}'$ is the probability measure in system i, and $\mathbb{E}$ is based on probability measure in the original system.

Proof For any outcome (sample path) $\omega \in C_i$, let $X(\omega)$ denotes the value of the random variable X on the sample path ω. Then we have

$$\begin{aligned}
\mathbb{P}(\{\omega\}) = &\prod_{k=1}^{K} \mathbb{P}\left(S_{k,1} = S_{k,1}(\omega)\right) \prod_{h=1}^{I_k(\omega)-1} \mathbb{P}\left(A_{k,h} = A_{k,h}(\omega) | S_{k,h}(\omega), \phi_{k,h}(\omega), \pi\right) \\
&\mathbb{P}\left(R_{k,h} = R_{k,h}(\omega) | A_{k,h}(\omega)\right) \left(1 - q\left(S_{k,h}(\omega), R_{k,h}(\omega)\right)\right) \\
&\mathbb{P}\left(A_{k,I_k(\omega)} = A_{k,I_k(\omega)}(\omega) | S_{k,I_k(\omega)}(\omega), \phi_{k,I_k(\omega)}(\omega), \pi\right) \\
&\mathbb{P}\left(R_{k,I_k(\omega)} = R_{k,I_k(\omega)}(\omega) | A_{k,I_k(\omega)}(\omega)\right) q\left(S_{k,I_k(\omega)}(\omega), R_{k,I_k(\omega)}(\omega)\right).
\end{aligned}$$

Let $k(s)$ and $h(s)$ denote the episode number and time step when a_i was pulled for the s-th time, respectively. Hence, we have

$$\begin{aligned}
\mathbb{P}'(\{\omega\}) = \mathbb{P}(\{\omega\}) \prod_{s=1}^{N_{T(K,\pi)+1}(a_i)(\omega)} &\left[\mathbb{1}\left\{R_{k(s)(\omega),h(s)(\omega)}(\omega) = 1\right\} \frac{\mu'(a_i)}{\mu(a_i)}\right. \\
&\left. + \mathbb{1}\left\{R_{k(s)(\omega),h(s)(\omega)}(\omega) = 0\right\} \frac{1-\mu'(a_i)}{1-\mu(a_i)}\right].
\end{aligned}$$

It then follows that

$$\begin{aligned}
\mathbb{P}'(C_i) = \sum_{\omega \in C_i} \mathbb{P}'(\{\omega\}) = \sum_{\omega \in C_i} \mathbb{P}(\{\omega\}) &\prod_{s=1}^{N_{T(K,\pi)+1}(a_i)(\omega)} \left[\mathbb{1}\left\{R_{k(s)(\omega),h(s)(\omega)}(\omega) = 1\right\} \frac{\mu'(a_i)}{\mu(a_i)}\right. \\
&\left. + \mathbb{1}\left\{R_{k(s)(\omega),h(s)(\omega)}(\omega) = 0\right\} \frac{1-\mu'(a_i)}{1-\mu(a_i)}\right] \\
= &\mathbb{E}\left[\mathbb{1}_{C_i} \prod_{s=1}^{N_{T(K,\pi)+1}(a_i)} \mathbb{1}\left\{R_{k(s),h(s)} = 1\right\} \frac{\mu'(a_i)}{\mu(a_i)} + \mathbb{1}\left\{R_{k(s),h(s)} = 0\right\} \frac{1-\mu'(a_i)}{1-\mu(a_i)}\right] \\
= &\mathbb{E}\left[\mathbb{1}_{C_i} \exp\left(-\hat{\mathrm{kl}}_{N_{T(K,\pi)+1}(a_i)}\right)\right],
\end{aligned}$$

where the last equality is by the definition of $\hat{\mathrm{kl}}_n$ in (95). ■

By Lemma 17 and $\hat{\mathrm{kl}}_{N_{T(K,\pi)+1}(a_i)} \leqslant (1 - \frac{\epsilon}{2}) \log K$ in the definition of C_i, we have

$$\mathbb{P}'(C_i) = \mathbb{E}\left[\mathbb{1}_{C_i} \exp\left(-\hat{\mathrm{kl}}_{N_{T(K,\pi)+1}(a_i)}\right)\right] \geqslant \mathbb{P}(C_i)\, K^{-\left(1-\frac{\epsilon}{2}\right)}.$$

It follows that

$$\mathbb{P}(C_i) \leqslant K^{\left(1-\frac{\epsilon}{2}\right)} \mathbb{P}'(C_i) \leqslant K^{\left(1-\frac{\epsilon}{2}\right)} \mathbb{P}'\left(N_{T(K,\pi)+1}(a_i) \leqslant \frac{1-\epsilon}{\mathrm{kl}(\mu(a_i), \mu'(a_i))} \log K\right)$$

$$
\begin{aligned}
=&K^{\left(1-\frac{\epsilon}{2}\right)}\mathbb{P}'\left(\sum_{j\neq i}N_{T(K,\pi)+1}(a_j)\geqslant T(K,\pi)-\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\right)\\
\leqslant&K^{\left(1-\frac{\epsilon}{2}\right)}\mathbb{P}'\left(\sum_{j\neq i}N_{T(K,\pi)+1}(a_j)\geqslant K-\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\right)\\
\leqslant&K^{\left(1-\frac{\epsilon}{2}\right)}\frac{\mathbb{E}'\left[\sum_{j\neq i}N_{T(K,\pi)+1}(a_j)\right]}{K-\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K},
\end{aligned}
\tag{97}
$$

where the second inequality is by the definition of C_i, the first equality is due to the fact that $\sum_j N_{T(K,\pi)+1}(a_j)=T(K,\pi)$, the third inequality is due to the fact that $T(K,\pi)\geqslant K$, and the last inequality is by Markov's inequality, where $\mathbb{E}'$ is based on probability measure in system i. Since $\pi\in\Pi_{\text{cons}}$, by the definition of consistent policies in Definition 5, we have

$$
\mathbb{E}'[\mathrm{Reg}_\pi(K)]=o\left(K^\alpha\right) \tag{98}
$$

for any $\alpha>0$. Similar to (93), for system i, we can obtain

$$
\begin{aligned}
\mathbb{E}'[\mathrm{Reg}_\pi(K)]\geqslant&\sum_{j\neq i}\mathbb{E}'\left[\sum_{t=1}^{T(K,\pi)}\mathbb{1}\{A_t=a_j\}\right]\left(V'^*(1)-Q'^*(1,a_j)\right)\\
\geqslant&\min_{j\neq i}\left(V'^*(1)-Q'^*(1,a_j)\right)\mathbb{E}'\left[\sum_{j\neq i}N_{T(K,\pi)+1}(a_j)\right],
\end{aligned}
\tag{99}
$$

where V'^* and Q'^* are the optimal value function and optimal Q function for system i, respectively. Since $\mu'(a_i)>\mu(a_j)$ for any $j\neq i$, similar to the original system, it can be verified that $\min_{j\neq i}\left(V'^*(1)-Q'^*(1,a_j)\right)>0$. Hence, from (99), we have

$$
\mathbb{E}'\left[\sum_{j\neq i}N_{T(K,\pi)+1}(a_j)\right]\leqslant\frac{\mathbb{E}'[\mathrm{Reg}_\pi(K)]}{\min_{j\neq i}\left(V'^*(1)-Q'^*(1,a_j)\right)}=o\left(K^\alpha\right)
$$

for any $\alpha>0$, where the equality is by (98). It then follows from (97) that

$$
\mathbb{P}\left(C_i\right)\leqslant K^{\left(1-\frac{\epsilon}{2}\right)}\frac{o\left(K^\alpha\right)}{K-\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K}
$$

for any $\alpha>0$. Let $\alpha=\frac{\epsilon}{4}$. Then we have

$$
\mathbb{P}\left(C_i\right)\leqslant K^{\left(1-\frac{\epsilon}{2}\right)}\frac{o\left(K^{\frac{\epsilon}{4}}\right)}{K-\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K}=o\left(K^{-\frac{\epsilon}{4}}\right). \tag{100}
$$

Note that by definition of C_i and the law of total probability, we have

$$
\mathbb{P}\left(N_{T(K,\pi)+1}(a_i)\leqslant\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\right)
$$

$$
\begin{aligned}
=&\mathbb{P}\left(C_i\right)+\mathbb{P}\left(N_{T(K,\pi)+1}(a_i)\leqslant\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K,\hat{\mathrm{kl}}_{N_{T(K,\pi)+1}(a_i)}>(1-\frac{\epsilon}{2})\log K\right)\\
\leqslant&\mathbb{P}\left(C_i\right)+\mathbb{P}\left(\hat{\mathrm{kl}}_{N_{T(K,\pi)+1}(a_i)}-N_{T(K,\pi)+1}(a_i)\mathrm{kl}(\mu(a_i),\mu'(a_i))>\frac{\epsilon}{2}\log K,\right.\\
&\qquad\left.N_{T(K,\pi)+1}(a_i)\leqslant\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\right)\\
\leqslant&\mathbb{P}\left(C_i\right)+\sum_{n=1}^{\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K}\mathbb{P}\left(\hat{\mathrm{kl}}_n-n\mathrm{kl}(\mu(a_i),\mu'(a_i))>\frac{\epsilon}{2}\log K\right)\\
\leqslant&\mathbb{P}\left(C_i\right)+\sum_{n=1}^{\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K}\exp\left(-\frac{\epsilon^2(\log K)^2}{2n\left|\log\frac{\mu(a_i)}{\mu'(a_i)}-\log\frac{1-\mu(a_i)}{1-\mu'(a_i)}\right|^2}\right)\\
\leqslant&\mathbb{P}\left(C_i\right)+\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\exp\left(-\frac{\epsilon^2(\log K)^2}{2\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\left|\log\frac{\mu(a_i)}{\mu'(a_i)}-\log\frac{1-\mu(a_i)}{1-\mu'(a_i)}\right|^2}\right)\\
=&\mathbb{P}\left(C_i\right)+\frac{(1-\epsilon)\log K}{\mathrm{kl}(\mu(a_i),\mu'(a_i))K^{\frac{\epsilon^2\mathrm{kl}(\mu(a_i),\mu'(a_i))}{2(1-\epsilon)\left|\log\frac{\mu(a_i)}{\mu'(a_i)}-\log\frac{1-\mu(a_i)}{1-\mu'(a_i)}\right|^2}}},
\end{aligned}
\tag{101}
$$

where the second inequality is by union bound over all possible values of $N_{T(K,\pi)+1}(a_i)$, and the third inequality is by Hoeffding's inequality. It follows from (100) and (101) that

$$
\lim_{K\to\infty}\mathbb{P}\left(N_{T(K,\pi)+1}(a_i)\leqslant\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\right)=0,
$$

which implies that

$$
\lim_{K\to\infty}\mathbb{P}\left(N_{T(K,\pi)+1}(a_i)>\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\right)=1. \tag{102}
$$

By Markov's inequality, we have

$$
\mathbb{P}\left(N_{T(K,\pi)+1}(a_i)>\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\right)\leqslant\frac{\mathbb{E}\left[N_{T(K,\pi)+1}(a_i)\right]}{\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K}.
$$

Therefore, we have

$$
\mathbb{E}\left[N_{T(K,\pi)+1}(a_i)\right]\geqslant\mathbb{P}\left(N_{T(K,\pi)+1}(a_i)>\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K\right)\frac{1-\epsilon}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\log K. \tag{103}
$$

Hence, it follows from (93) and (103) that

$$
\mathbb{E}[\mathrm{Reg}_\pi(K)]\geqslant\sum_{i=2}^{M}\mathbb{E}\left[N_{T(K,\pi)+1}(a_i)\right](V^*(1)-Q^*(1,a_i))
$$

$P(s'\|s,a)$		Next state s'		
		$(1-\theta)x$	$(1-\theta)x+\theta$	g
State s	x	$(1-\mu(a))[1-q((1-\theta)x)]$	$\mu(a)[1-q((1-\theta)x+\theta)]$	$(1-\mu(a))q((1-\theta)x)+\mu(a)q((1-\theta)x+\theta)$
	g	0	0	1

Table 2: Transition probabilities $P(s'|s,a)$

$$\begin{aligned}
&\geqslant \sum_{i=2}^{M} \mathbb{P}\left(N_{T(K,\pi)+1}(a_i) > \frac{(1-\epsilon)\log K}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\right) \frac{(1-\epsilon)\log K}{\mathrm{kl}(\mu(a_i),\mu'(a_i))} \left(V^*(1) - Q^*(1,a_i)\right) \\
&\geqslant \sum_{i=2}^{M} \mathbb{P}\left(N_{T(K,\pi)+1}(a_i) > \frac{(1-\epsilon)\log K}{\mathrm{kl}(\mu(a_i),\mu'(a_i))}\right) \frac{(1-\epsilon)\log K}{(1+\epsilon)\mathrm{kl}(\mu(a_i),\mu(a_1))} \left(V^*(1) - Q^*(1,a_i)\right),
\end{aligned} \tag{104}$$

where the last inequality follows from (94). From (102) and (104), we have

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_\pi(K)]}{\log K} \geqslant \sum_{i\neq 1} \frac{1-\epsilon}{(1+\epsilon)\mathrm{kl}(\mu(a_i),\mu(a_1))} \left(V^*(1) - Q^*(1,a_i)\right).$$

Appendix C. Extension to the General-State Setting

In this section, we present the details of the state transition, the details of the proofs, and the simulation results in the general-state setting.

C.1 State Transition

The transition probabilities $P(s'|s,a)$ while pulling arm a are shown in Table 2, where $x \in [0,1]$.

C.2 Proof of Lemma 10

If the model is known, this problem can be viewed as a SSP problem (Bertsekas and Tsitsiklis, 1991). Since $q(s) > 0$ for any $s \in [0,1]$, all policies are proper. Hence, by the results of Bertsekas and Tsitsiklis (1991), there exists a stationary optimal policy. Therefore, it is enough to consider only stationary policies for π^*. Define the optimal value function V^* and optimal Q function Q^* the same way as (22) and (23). Then for $s \neq g$ and $a \in \mathcal{A}$, we have the Bellman equation as follows

$$\begin{aligned}
V^*(s) =& \max_a Q^*(s,a) \\
Q^*(s,a) =& \mu(a) + (1-\mu(a))\left[1-q((1-\theta)s)\right]V^*((1-\theta)s) \\
&+ \mu(a)\left[1-q((1-\theta)s+\theta)\right]V^*((1-\theta)s+\theta).
\end{aligned}$$

Hence, for any $s_1, s_2 \in [0,1]$ such that $s_1 \geqslant s_2$ and any $a \in \mathcal{A}$, we have

$$\begin{aligned}
&Q^*(s_1,a) - Q^*(s_2,a) \\
=&(1-\mu(a))\Big\{\left[1-q((1-\theta)s_1)\right]V^*((1-\theta)s_1) - \left[1-q((1-\theta)s_2)\right]V^*((1-\theta)s_2)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&+\mu(a)\Big\{\left[1-q((1-\theta)s_1+\theta)\right]V^*((1-\theta)s_1)+\theta)\\
&\qquad-\left[1-q((1-\theta)s_2+\theta)\right]V^*((1-\theta)s_2+\theta)\Big\}\\
\geqslant&(1-\mu(a))\left[1-q((1-\theta)s_1)\right]\left[V^*((1-\theta)s_1)-V^*((1-\theta)s_2)\right]\\
&+\mu(a)\left[1-q((1-\theta)s_1+\theta)\right]\left[V^*((1-\theta)s_1+\theta)-V^*((1-\theta)s_2+\theta)\right],
\end{aligned}
$$

where the inequality is by Assumption 2. Therefore, we have

$$
\begin{aligned}
&V^*(s_1)-V^*(s_2)=\max_a Q^*(s_1,a)-\max_a Q^*(s_2,a)=\max_a Q^*(s_1,a)-Q^*(s_2,a')\\
\geqslant&Q^*(s_1,a')-Q^*(s_2,a')\\
\geqslant&(1-\mu(a))\left[1-q((1-\theta)s_1)\right]\left[V^*((1-\theta)s_1)-V^*((1-\theta)s_2)\right]\\
&+\mu(a)\left[1-q((1-\theta)s_1+\theta)\right]\left[V^*((1-\theta)s_1+\theta)-V^*((1-\theta)s_2+\theta)\right],
\end{aligned}
\tag{105}
$$

where $a' := \text{argmax}_a\, Q^*(s_2,a)$. Note that $1-q(s)<1$ for any $s\in[0,1]$ by Assumption 2. Hence, by iteratively applying (105), we can obtain $V^*(s_1)-V^*(s_2)\geqslant 0$, which means that $V^*(s)$ is non-decreasing in s. Hence, for any $s\in[0,1]$ and any $i\in\{2,...,M\}$, we have

$$
\begin{aligned}
&Q^*(s,a_1)-Q^*(s,a_i)\\
=&(\mu(a_1)-\mu(a_i))+(\mu(a_1)-\mu(a_i))\Big\{[1-q((1-\theta)s+\theta)]V^*((1-\theta)s+\theta)\\
&\qquad-[1-q((1-\theta)s)]V^*((1-\theta)s)\Big\}\geqslant 0,
\end{aligned}
\tag{106}
$$

where the inequality is by Assumption 2 and the monotonicity of V^*. Therefore, the genie-aided optimal policy is always pulling Arm a_1.

C.3 Some Examples of the Abandonment Probability Functions

We present some examples of the abandonment probability functions $q(\cdot)$ that satisfy $V^*(s_1)-Q^*(s_1,a)\leqslant V^*(s_2)-Q^*(s_2,a)$ for any $a\in\mathcal{A}, s_1,s_2\in[0,1], s_1\geqslant s_2$.

For any $a\in\mathcal{A}, s_1,s_2\in[0,1], s_1\geqslant s_2$, we have

$$
\begin{aligned}
&[V^*(s_1)-Q^*(s_1,a)]-[V^*(s_2)-Q^*(s_2,a)]\\
=&[Q^*(s_1,a_1)-Q^*(s_1,a)]-[Q^*(s_2,a_1)-Q^*(s_2,a)]\\
=&(\mu(a_1)-\mu(a))\Big\{[1-q((1-\theta)s_1+\theta)]V^*((1-\theta)s_1+\theta)-[1-q((1-\theta)s_1)]V^*((1-\theta)s_1)\\
&\qquad-[1-q((1-\theta)s_2+\theta)]V^*((1-\theta)s_2+\theta)+[1-q((1-\theta)s_2)]V^*((1-\theta)s_2)\Big\},
\end{aligned}
\tag{107}
$$

where the first quality is by Lemma 10, and the second equality is by (106). From (107), we know that the sign of $[V^*(s_1)-Q^*(s_1,a)]-[V^*(s_2)-Q^*(s_2,a)]$ depends only on the abandonment probability function $q(\cdot)$ and $V^*(\cdot)$. By Lemma 10, the Bellman equation of $V^*(s)$ is as follows

$$
V^*(s)=\mu(a_1)+(1-\mu(a_1))\left[1-q((1-\theta)s)\right]V^*((1-\theta)s)
$$

$$+ \mu(a_1)\left[1 - q((1-\theta)s + \theta)\right] V^*((1-\theta)s + \theta)$$

Therefore, if we know the abandonment function $q(\cdot)$ and $\mu(a_1)$, we can numerically calculate $V^*(\cdot)$ and then determine the sign of $[V^*(s_1) - Q^*(s_1, a)] - [V^*(s_2) - Q^*(s_2, a)]$ by (106). For example, let $\mu(a_1) = 0.9$, $\theta = 0.5$, and

$$q(s) = 1 - \frac{\log(c_6 s + 1)}{\log(c_6 + 1)}, \tag{108}$$

where c_6 is a constant. Figure 4 shows the abandonment probability functions (108) for $c_6 = 5$, $c_6 = 50$, and $c_6 = 1000$. We numerically check that for $c_6 = 5$, $c_6 = 50$, and $c_6 = 1000$,

$$[V^*(s_1) - Q^*(s_1, a)] \leqslant [V^*(s_2) - Q^*(s_2, a)] \tag{109}$$

for any $a \in \mathcal{A}, s_1, s_2 \in [0, 1], s_1 \geqslant s_2$. We conjecture that (109) holds for any $c_6 \geqslant 5$.

C.4 Details of CONT-ULCB and CONT-KL-ULCB Algorithms

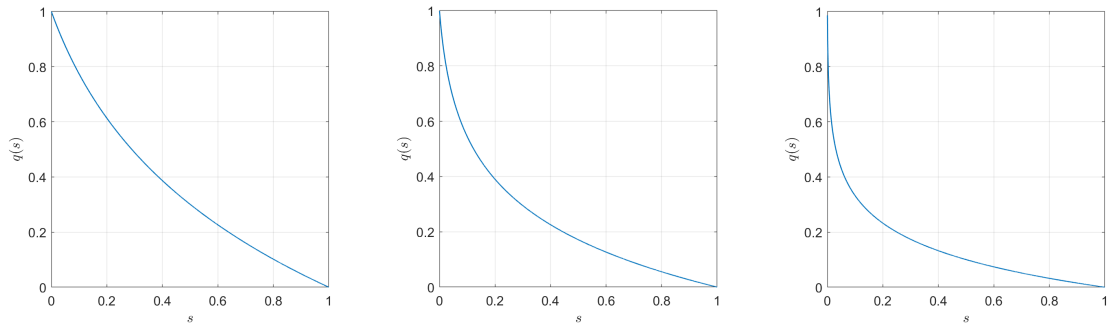
The CONT-ULCB algorithm is shown in Algorithm 2. The CONT-KL-ULCB algorithm replaces $\tilde{\mu}_t^{S_{k,h}}(a)$ in Algorithm 2 with KL divergence, i.e., for all $a \in \mathcal{A}$,

$$\tilde{\mu}_t^{S_{k,h}}(a) = \begin{cases} \min\left\{p : \mathrm{kl}(\bar{\mu}_t(a), p) N_t(a) \leqslant (1 - 2S_{k,h}) \log t + c \log(\log t)\right\}, & S_{k,h} \leqslant \frac{1}{2} \\ \max\left\{p : \mathrm{kl}(\bar{\mu}_t(a), p) N_t(a) \leqslant (2S_{k,h} - 1) \log t + c \log(\log t)\right\}, & S_{k,h} > \frac{1}{2}. \end{cases}$$

C.5 Simulation Results for the General-State Setting

Consider the MAB-A problem of the general-state setting. Let the abandonment probability function $q(\cdot)$ be

$$q(s) = 1 - \frac{\log(c_6 s + 1)}{\log(c_6 + 1)}$$



(a) $c_6 = 5$ (b) $c_6 = 50$ (c) $c_6 = 1000$

Figure 4: Examples of abandonment probability functions.

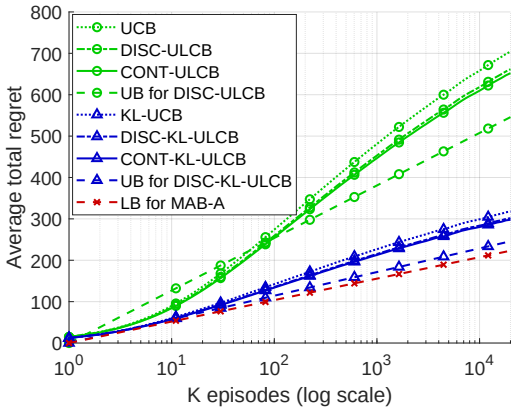
Algorithm 2 CONT-ULCB Algorithm

Initialize: $N_1(a) \leftarrow 0$, $\bar{\mu}_1(a) \leftarrow 0$ for all $a \in \mathcal{A}$, $t \leftarrow 1$, c_0, c_1, c.
for episode $k = 1, ..., K$ **do**
  $h \leftarrow 1$, $S_{k,1} \leftarrow$ initial state of episode k, $S_{k,1} \in [0, 1]$
  while $S_{k,h} \neq g$ **do**
    if there exists Arm a' such that $N_t(a') = 0$ **then**
      play Arm $A_{k,h} = a'$ and observe $R_{k,h}$
    else
      Let $\tilde{\mu}_t^{S_{k,h}}(a) = \bar{\mu}_t(a) + (2S_{k,h} - 1)\sqrt{\frac{\log t + c \log(\log t)}{2N_t(a)}}$ for all $a \in \mathcal{A}$
      Take the action $A_{k,h} \in \text{argmax}_a \tilde{\mu}_t^{S_{k,h}}(a)$ and observe $R_{k,h}$.
    end if
    if abandonment occurs **then** $S_{k,h+1} = g$
    else $S_{k,h+1} = (1 - \theta)S_{k,h} + \theta R_{k,h}$
    end if
    Define $(S_t, A_t, S'_t, R_t) := (S_{k,h}, A_{k,h}, S_{k,h+1}, R_{k,h})$
    Update: $N_{t+1}(A_t) = N_t(A_t) + 1$ and $N_{t+1}(a) = N_t(a)$ $\forall a \neq A_t$
    Update: $\bar{\mu}_{t+1}(A_t) = \frac{\bar{\mu}_t(A_t)N_t(A_t)+R_t}{N_{t+1}(A_t)}$ and $\bar{\mu}_{t+1}(a) = \bar{\mu}_t(a)$ $\forall a \neq A_t$
    $t \leftarrow t + 1$, $h \leftarrow h + 1$
  end while
end for

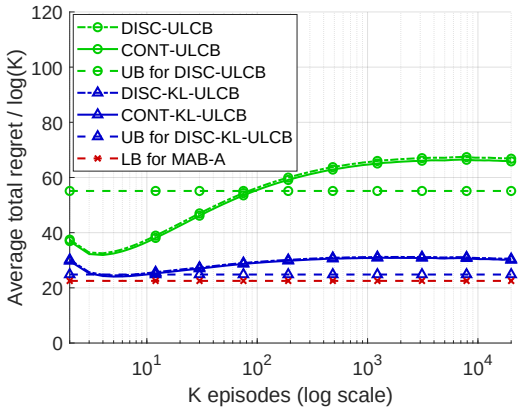
for any $s \in [0, 1]$, where c_6 is a constant. Let the forgetting factor $\theta = 0.5$ in the simulation. We present the simulation results for our proposed DISC-ULCB, CONT-ULCB, DISC-KL-ULCB, and CONT-KL-ULCB algorithms. Let $n = 4$ for the discretization of DISC-ULCB and DISC-KL-ULCB. We simulated 2×10^4 episodes with 10^7 independent runs. Simulation results are shown in Figure 5, Figure 6, and Figure 7 for different sets of arms and different abandonment probabilities (different c_6).

Remark 18 *For Figure 5(a), the* 95% *confidence bounds are at most* ± 4.43. *For Figure 6(a), the* 95% *confidence bounds are at most* ± 3.11. *For Figure 7(a), the* 95% *confidence bounds are at most* ± 0.12.

From Figure 5(a), Figure 6(a), and Figure 7(a), we can see in all the three different settings that both DISC-ULCB and CONT-ULCB algorithms outperform the traditional UCB in terms of average cumulative regret, and that both DISC-KL-ULCB and CONT-KL-ULCB algorithms outperform the traditional KL-UCB. Moreover, CONT-ULCB and CONT-KL-ULCB perform slightly better than DISC-ULCB and DISC-KL-ULCB, respecively. These figures also show that the slopes of DISC-ULCB and DISC-KL-ULCB converges to the slopes of their corresponding upper bounds. Also, in Figure 5(b), Figure 6(b), and Figure 7(b) where the Y-axis is the average cumulative regret divided by $\log K$, the curves of DISC-ULCB and DISC-KL-ULCB go towards their corresponding asymptotic upper bounds, which confirms our theoretical results.

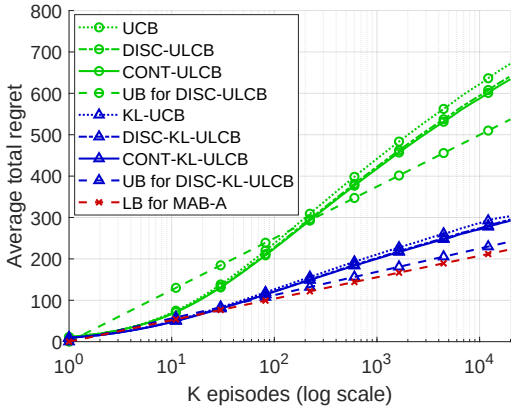


(a) Comparison among algorithms

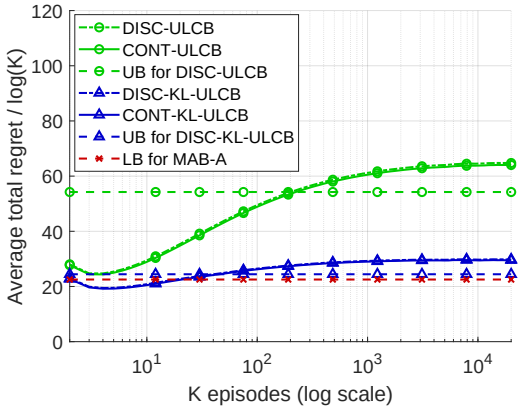


(b) Upper bound (UB) and lower bound (LB)

Figure 5: Simulation results for the general-state setting, $M = 2$, $\mu(a_1) = 0.9$, $\mu(a_2) = 0.8$, $c_6 = 1000$.

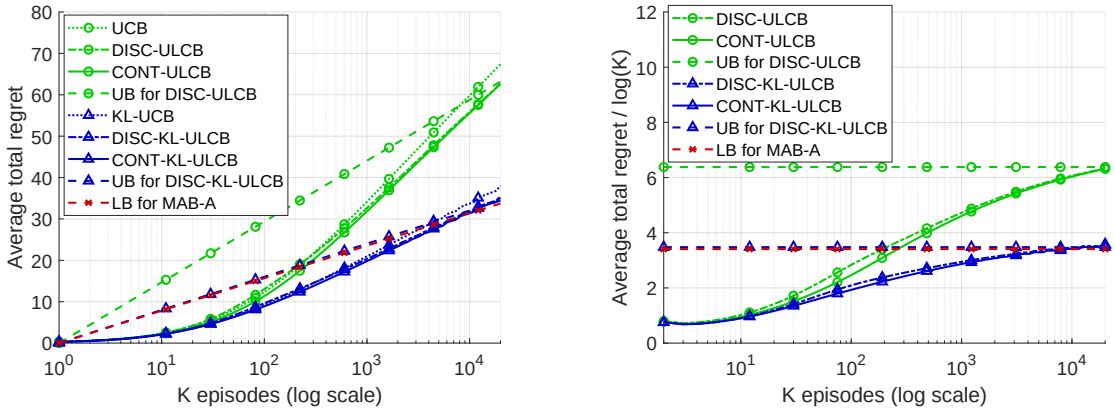


(a) Comparison among algorithms



(b) Upper bound (UB) and lower bound (LB)

Figure 6: Simulation results for the general-state setting, $M = 2$, $\mu(a_1) = 0.9$, $\mu(a_2) = 0.8$, $c_6 = 100$.



(a) Comparison among algorithms

(b) Upper bound (UB) and lower bound (LB)

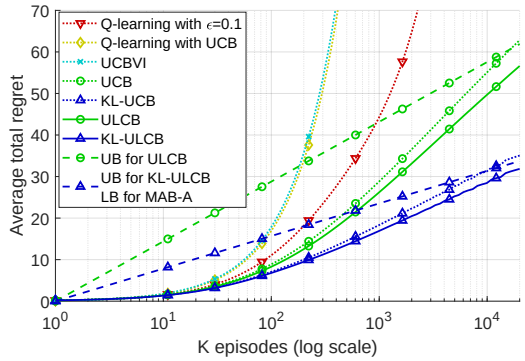
Figure 7: Simulation results for the general-state setting, $M = 2$, $\mu(a_1) = 0.2$, $\mu(a_2) = 0.1$, $c_6 = 1000$.

Appendix D. Simulation Parameters and Additional Simulations

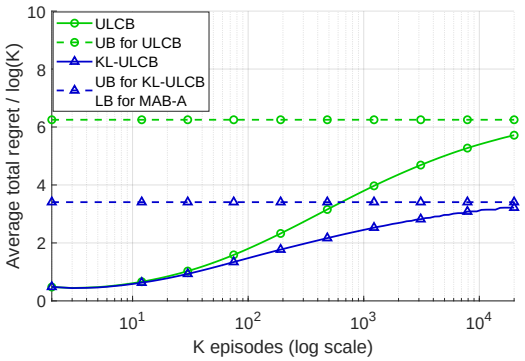
In the simulation of Q-learning with ϵ-greedy, we set $\epsilon = 0.1$. The learning rate is set to be $\frac{1}{N(s,a)}$, where $N(s,a)$ is the number of times the state-action pair (s,a) has been visited. For Q-learning with UCB, we set the episode length parameter H to be the maximum expected episode length in MAB-A, which is the expected episode length under the policy of always pulling the optimal arm a_1. The number of episodes K is set to be 1000. The constant c in the bonus term is set to be 4. For UCBVI, we also set the episode length parameter H to be the expected episode length under the best policy and the number of episodes K is set to be 1000. The probability parameter δ is set to be 0.001. The bonus term is modified to $b(s,a) = 7H\log(5SAHK/\delta)\sqrt{\frac{1}{\min\{N(a),N_{\mathrm{r}}(s,1),N_{\mathrm{r}}(s,0)\}}}$, where $N(a)$ is the number of pulls of arm a and $N_{\mathrm{r}}(s,r)$ is the number of visits of state-reward pair (s,r). This is natural because $N(a)$, $N_{\mathrm{r}}(s,1)$, and $N_{\mathrm{r}}(s,0)$ reflect the uncertainty of the estimates of the mean rewards and the abandonment probabilities.

We did additional simulations for different sets of arms and different abandonment probabilities, as shown in Figure 8 and Figure 9. The other settings are the same as those in Section 5. The same conclusion holds, i.e., our proposed ULCB and KL-ULCB outperform the traditional UCB and KL-UCB, respectively, our algorithms have order-wise lower regret than Q-learning with ϵ-greedy, Q-learning with UCB, and UCBVI, and the simulation results are also consistent with our theoretical results.

In order to further understand the exploration and the exploitation in MAB-A problem, we run ULCB algorithms with different c_0 and c_1, which are the state-dependent exploration-exploitation coefficients in the indices for state 0 and state 1, respectively. Larger coefficient means more exploration. The results are shown in Figure 10. In Figure 10(a), c_0 is fixed, and as c_1 increases, the cumulative regret decreases. In Figure 10(b), c_1 is fixed, and as c_0 increases, the cumulative regret increases. These changes of regret are consistent with our theoretical results which suggest more exploration in state 1.

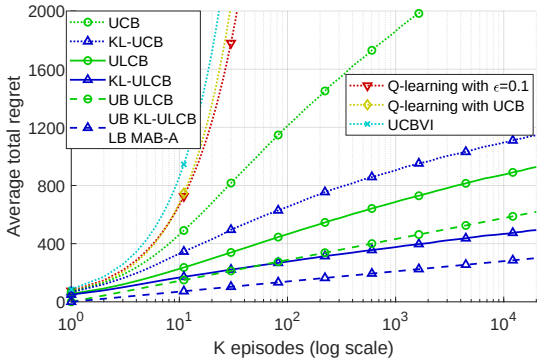


(a) Comparison among algorithms

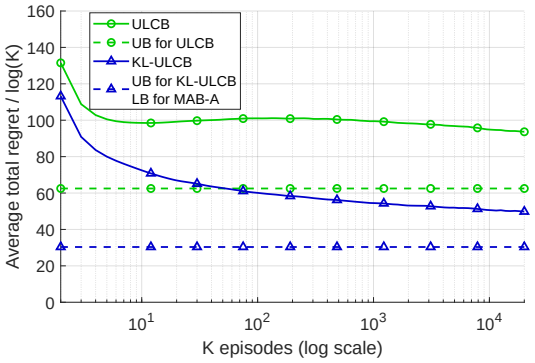


(b) Upper bound (UB) and lower bound (LB)

Figure 8: Simulation results, $M = 2$, $\mu(a_1) = 0.2$, $\mu(a_2) = 0.1$, $q(0,0) = 1$, $q(1,0) = q(0,1) = q(1,1) = 0$.

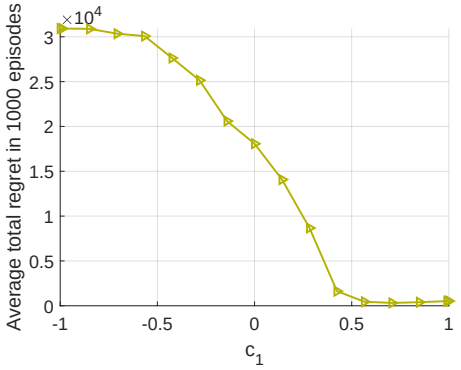


(a) Comparison among algorithms

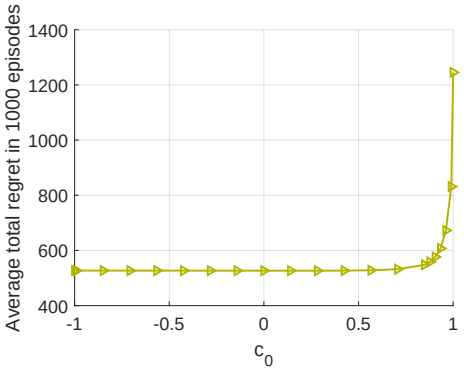


(b) Upper bound (UB) and lower bound (LB)

Figure 9: Simulation results, $M = 3$, $\mu(a_1) = 0.9$, $\mu(a_2) = 0.8$, $\mu(a_3) = 0.5$, $q(0,0) = 1$, $q(1,0) = q(0,1) = q(1,1) = 0$.



(a) Fix $c_0 = -1$



(b) Fix $c_1 = 1$

Figure 10: ULCB: different coefficients c_0 and c_1, $q(0,0) = 1$, $q(1,1) = q(1,0) = q(0,1) = 0$.

Appendix E. Results for the Opposite Case—(8) Does Not Hold

Note that you can see in (29) in Appendix B.2 that the inequality $V^*(1) - Q^*(1,a) \geqslant V^*(0) - Q^*(0,a)$ or $V^*(1) - Q^*(1,a) \leqslant V^*(0) - Q^*(0,a)$ does not depend on a. Hence, at least one of the two cases holds. In this section, we consider the case where $V^*(1) - Q^*(1,a) \geqslant V^*(0) - Q^*(0,a)$, i.e., condition (8) does not hold. One example is that $q(0,0) = 0.6, q(1,0) = q(0,1) = 0.5, q(1,1) = 0.1$. With modified algorithms, we have results similar to our main results.

E.1 Upper Bound for Modified ULCB

Theorem 19 *Let Assumption 1 hold. Suppose $V^*(1) - Q^*(1,a) \geqslant V^*(0) - Q^*(0,a)$ for any $a \neq a_1$. Then using Algorithm 1 with $c_0 = 1$, $c_1 = -1$, and $c = 4$, we have*

$$\limsup_{K\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_\pi(K)]}{\log K} \leqslant \sum_{i\neq 1} \frac{V^*(0) - Q^*(0,a_i)}{2(\mu(a_1) - \mu(a_i))^2}.$$

Note that in Theorem 2 we use $c_0 = -1$ and $c_1 = 1$ while in Theorem 19 we use the opposite, i.e., $c_0 = 1$ and $c_1 = -1$. The proof is symmetric to that of Theorem 2 and hence is omitted.

E.2 Upper Bound for Modified KL-ULCB

For the case where $V^*(1) - Q^*(1,a) \geqslant V^*(0) - Q^*(0,a)$, we use a modified KL-ULCB algorithm, which replaces the indices $\tilde{\mu}_t^0(a)$ and $\tilde{\mu}_t^1(a)$ in (4) and (5) with

$$\tilde{\mu}_t^0(a) = \max\{p : \mathrm{kl}(\bar{\mu}_t(a), p) N_t(a) \leqslant c_0 \log t + c \log(\log t)\}$$
$$\tilde{\mu}_t^1(a) = \min\{p : \mathrm{kl}(\bar{\mu}_t(a), p) N_t(a) \leqslant c_1 \log t + c \log(\log t)\}.$$

Theorem 20 *Let all the assumptions in Theorem 19 hold. Then using the above modified KL-ULCB algorithm with $c_0 = c_1 = 1$ and $c = 4$, we have*

$$\limsup_{K\to\infty} \frac{\mathbb{E}[\mathrm{Reg}_\pi(K)]}{\log K} \leqslant \sum_{i\neq 1} \frac{V^*(0) - Q^*(0,a_i)}{\mathrm{kl}(\mu(a_i), \mu(a_1))}. \tag{110}$$

The proof is symmetric to that of Theorem 4 and hence is omitted.

E.3 Lower Bound for the Opposite Case

Theorem 21 *Let all the assumptions in Theorem 19 hold. Let π be a consistent policy, i.e., $\pi \in \Pi_{\text{cons}}$. Then for any $\mu(a_1), ..., \mu(a_M), q(0,0), q(0,1), q(1,0), q(1,1)$ satisfying the assumptions, we have*

$$\liminf_{K\to\infty} \frac{\mathbb{E}[\text{Reg}_\pi(K)]}{\log K} \geqslant \sum_{i\neq 1} \frac{V^*(0) - Q^*(0,a_i)}{\text{kl}(\mu(a_i), \mu(a_1))}. \tag{111}$$

The proof is symmetric to that of Theorem 6 and hence is omitted. From Theorem 20 and Theorem 21, the upper bound obtained by the modified KL-ULCB algorithm matches the lower bound asymptotically for the case where $V^*(1) - Q^*(1,a) \geqslant V^*(0) - Q^*(0,a)$ for any $a \neq a_1$.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Omer Ben-Porat, Lee Cohen, Liu Leqi, Zachary C Lipton, and Yishay Mansour. Modeling attrition in recommender systems with departing bandits. *arXiv preprint arXiv:2203.13423*, 2022.

Dimitri P Bertsekas and John N Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

Qingpeng Cai, Shuchang Liu, Xueliang Wang, Tianyou Zuo, Wentao Xie, Bin Yang, Dong Zheng, Peng Jiang, and Kun Gai. Reinforcing user retention in a billion scale short video recommender system. *arXiv preprint arXiv:2302.01724*, 2023.

Junyu Cao and Wei Sun. Dynamic learning of sequential choice bandit problem under marketing fatigue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 3264–3271, 2019.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. In *Neural Information Processing Systems*, 2021.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. In *Neural Information Processing Systems*, 2021.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pages 267–274, 2002.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jay Mardia, Jiantao Jiao, Ervin Tánczos, Robert D Nowak, and Tsachy Weissman. Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*, 9(4):813–850, 2020.

Sophia Petrillo. What makes tiktok so addictive?: An analysis of the mechanisms underlying the world's latest social media craze. *Brown Undergraduate Journal of Public Health*, 2021.

Sahar F Sabbeh. Machine-learning techniques for customer retention: A comparative study. *International Journal of advanced computer Science and applications*, 9(2), 2018.

Sven Schmit and Ramesh Johari. Learning with abandonment. In *International Conference on Machine Learning*, pages 4509–4517. PMLR, 2018.

Jean Tarbouriech, Runlong Zhou, Simon S. Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. In *Neural Information Processing Systems*, 2021.

Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2016.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Regret bounds for stochastic shortest path problems with linear function approximation. *arXiv preprint arXiv:2105.01593*, 2021.

Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, United Kingdom, 1989.

Huasen Wu, Xueying Guo, and Xin Liu. Adaptive exploration-exploitation tradeoff for opportunistic bandits. In *International Conference on Machine Learning*, pages 5306–5314. PMLR, 2018.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR, 2021.