HOKKAIDO UNIVERSITY

Title	KL-UCB-Based Policy for Budgeted Multi-Armed Bandits with Stochastic Action Costs
Author(s)	WATANABE, Ryo; KOMIYAMA, Junpei; NAKAMURA, Atsuyoshi et al.
Citation	IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, E100.A(11), 2470-2486 https://doi.org/10.1587/transfun.E100.A.2470
Issue Date	2017-11-01
Doc URL	https://hdl.handle.net/2115/89513
Rights	copyright©2017 IEICE
Type	journal article
File Information	IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 100-A_11_2470-2486.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

IEICE TRANS. FUNDAMENTALS, VOL.E100–A, NO.11 NOVEMBER 2017

PAPER

KL-UCB-Based Policy for Budgeted Multi-Armed Bandits with Stochastic Action Costs

Ryo WATANABE[†a)], **Junpei KOMIYAMA**[††], ***Nonmembers***, **Atsuyoshi NAKAMURA**[†], ***and*** **Mineichi KUDO**[†], ***Members***

SUMMARY We study the budgeted multi-armed bandit problem with stochastic action costs. In this problem, a player not only receives a reward but also pays a cost for an action of his/her choice. The goal of the player is to maximize the cumulative reward he/she receives before the total cost exceeds the budget. In the classical multi-armed bandit problem, a policy called KL-UCB is known to perform well. We propose KL-UCB-SC, an extension of this policy for the budgeted bandit problem. We prove that KL-UCB-SC is asymptotically optimal for the case of Bernoulli costs and rewards. To the best of our knowledge, this is the first result that shows asymptotic optimality in the study of the budgeted bandit problem. In fact, our regret upper bound is at least four times better than that of BTS, the best known upper bound for the budgeted bandit problem. Moreover, an empirical simulation we conducted shows that the performance of a tuned variant of KL-UCB-SC is comparable to that of state-of-the-art policies such as PD-BwK and BTS.

key words: *budgeted multi-armed bandits, asymptotically optimal policy, regret analysis*

1. Introduction

The *multi-armed bandit problem* is a classical problem studied by Thompson [1] and Robbins [2], which is formalized as a problem on the following repeated game. At each round, a player chooses one of the available actions and receives a corresponding reward without obtaining any information on the rewards of the other actions. The player's goal is to maximize his/her cumulative (total) reward over the all rounds. To achieve this goal, the player must balance *exploration* and *exploitation*; that is, the player must trade-off the value of choosing an action with currently highest average reward for large *immediate* gain, with the value of choosing an action with potentially highest average reward for large *future* gain.

In this Internet era, the multi-armed bandit has become a more popular study area because many online decision making tasks such as recommendation and online advertising can be formalized as multi-armed bandit problems. In order to apply to a wider range of real-world problems, many extensions and generalizations of the problem have been suggested. Among them, *budgeted multi-armed bandits* [3] is an extension in which the player is charged for the chosen action and can repeatedly choose actions until the total cost exceeds the given budget. Real-world problems that can be formalized as this problem are online selection of a rental ad space [3] and action decisions of battery-driven wireless sensor devices [4]. In the former example, an action and a cost corresponds to an ad slot and a charge of the slot, respectively. In the latter case, an action corresponds to sampling and data forwarding, and the cost corresponds to the energy that is consumed for each action. Moreover, Ding et al. considered the case of stochastic costs in this extension [5]. In this formulation, the costs can vary among rounds, which allows us to deal with an even broader range of applications, such as an online bidding optimization in sponsored search [6], [7], in which an action is a possible price to bid for ad space in the result page of a keyword search, and the second price must be paid when the advertiser is the winner of the auction. In this paper, *budgeted multi-armed bandits* will refer to this stochastic-cost version of the problem.

As is the classical multi-armed bandits, the performance of a policy in the budgeted multi-armed bandits is measured by a quantity called (pseudo) *regret*, which is the difference between the expected cumulative reward of the policy and the maximum expected cumulative reward among all policies. A policy is called optimal when its regret upper bound matches a known regret lower bound. Although several of the proposed policies are introduced with corresponding regret upper bounds, no analysis on the regret lower bound has been shown so far. Therefore, whether these policies are optimal or not is unknown.

The difficulty with budgeted multi-armed bandits lies in that both the rewards and the costs are stochastic; thus, one needs to take the uncertainty of both into consideration. A naïve UCB-type approach is to build an upper confidence bound by using an upper confidence bound of the reward and a lower confidence bound of the cost. Indeed, some policies, such as UCB-BV family [5] and PD-BwK [8], take this approach. We argue that such a two-bound approach can lead to a loose evaluation of the confidence bound. Instead, if one can combine the uncertainty of the reward and the cost into a single confidence bound, a tighter evaluation of the uncertainty is available, which yields a policy of better performance. Taking the above into consideration, we propose a deterministic policy KL-UCB-SC for the budgeted multi-armed bandits based on the KL-UCB [9]–[11] policy for classical multi-armed bandit problem. We build a natural extension of the KL-UCB index to the budgeted multi-armed

Manuscript received June 16, 2017.

[†]The authors are with Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.

[††]The author is with Institute of Industrial Science, the University of Tokyo, Tokyo, 153-8505 Japan.

a) E-mail: ryo@main.ist.hokudai.ac.jp

DOI: 10.1587/transfun.E100.A.2470

Copyright © 2017 The Institute of Electronics, Information and Communication Engineers

bandits and show its effectiveness for the budgeted bandit problem.

1.1 Contributions

The contributions of this paper can be summarized into the following four aspects:

- We propose the KL-UCB-SC policy, which is a natural extension of KL-UCB for the budgeted multi-armed bandits.
- Let the optimal action be the action of the best reward per cost in expectation, and let the suboptimal actions be the others. We give an upper bound of the expected number of selections for each suboptimal action. We also give a lower bound in the case of Bernoulli rewards and costs, whose leading term coefficient asymptotically coincides with that of the upper bound.
- In the variable cost setting, in which the number of the rounds $T(B)$ varies among runs based on the total cost B, we propose a convenient way to convert bounds on the number of selections in T rounds for fixed T into a regret for the total cost B and derive a regret upper bound of KL-UCB-SC and a regret lower bound for the problem in the case of Bernoulli rewards and costs. The regret upper bound holds for any reward and cost distribution bounded in $[0, 1]$ and is asymptotically optimal in the case of Bernoulli rewards and costs.
- We assess the performance of the proposed policy in a simulation and show that the performance of KL-UCB-SC+, a tuned variant, is comparable to that of BTS [12].

2. Related Work

Most of the recent studies on multi-armed bandits in the machine learning community can be categorized into stochastic and adversarial settings. The adversarial setting assumes no reward distribution, and rewards can be generated by an adversary who can adapt to the player's policy [13]. Therefore, the adversarial setting is suitable to model decision making in multi-player games. The stochastic setting, in which the reward on each action is assumed to belong to some stochastic process, yields policies of generally better performance when the distributions of rewards do not change very rapidly (e.g., recommendation for quasi-static population). In this paper, we focus on the latter setting.

The UCB (by Auer et al. [14], originally called as UCB1), is probably the most well-known policy for the stochastic multi-armed bandit problem. They also proved that UCB achieves a logarithmic expected regret [14]. UCB is not asymptotically optimal in the sense that the leading logarithmic term is not the best. Today, several asymptotically optimal policies are known, such as KL-UCB (Kullback-Leibler UCB) [9], [10], Thompson sampling and DMED [15]. Among them, the KL-UCB policy, the original idea of which also appeared in old papers [11], [16], is closest to the UCB in the sense that it explicitly uses the upper confidence bound. Thompson sampling [1] is an old technique based on the idea of posterior sampling, and its asymptotic optimality is proven by [17] recently. DMED is another algorithm that explicitly computes the likelihood of its action to be optimal, and explores actions that are candidates of the optimal.

In particular, this paper focuses on the budgeted bandit problem. Unlike the classical bandit problem where the action of the largest expected reward is of interest, in the budgeted bandit problem, the action of the largest expected reward per cost is sought. As for the budgeted multi-armed bandits, KUBE [4], a naïve extension of the UCB, is the first consistent policy that solves the bandit problem in the case of constant (but action-dependent) costs. In this case, the true optimal policy is to choose actions following to the solution of a corresponding unbounded knapsack problem. Although KUBE does not solve unbounded knapsack problems with estimated mean rewards, the strong consistency (that is, its expected regret is upper bounded subpolynomially to the budget B) is proved. [5] extended the problem so as to allow the action costs to be distributed. UCB-BV family [5] and PD-BwK [8] are policies of the naïve "two-bound approach" for this problem as described in Sect. 1. BTS [12], which is based on Thompson sampling, is the currently best policy for this problem. Our policy improves its regret upper bound by a factor of four and we show that our bound is the best possible.

3. Problem Settings

The budgeted multi-armed bandits [5] is described as a one-player game in which a player repeatedly chooses one of K actions until he/she uses up a given budget. At each round $t = 1, 2, \ldots$, a player chooses an action $I(t) \in [K] := \{1, \ldots, K\}$, and then he/she pays a cost $C_{I(t)}(t) \in [0, 1]$ and receives a reward $X_{I(t)}(t) \in [0, 1]$. The rewards $X_i(1), X_i(2), \ldots$ and costs $C_i(1), C_i(2), \ldots$ of an action i are i.i.d. random variables and are drawn from corresponding distributions with means μ_i and τ_i. The realizations $X_i(t)$ and $C_i(t)$ of unchosen actions $i \neq I(t)$ are not revealed to the player. We assume that $\mu_i, \tau_i \in (0, 1)$. Let $\tau_{\min}$ and $\tau_{\max}$ denote $\min\{\tau_i \mid i \in [K]\}$ and $\max\{\tau_i \mid i \in [K]\}$, respectively. Let $c_t = \sum_{s=1}^{t} C_{I(s)}(s)$ be the total cost. The player immediately stops choosing actions at the round when the total cost c_t exceeds the budget B. The game is summarized in the form of the pseudo-code in Algorithm 1.

Let us study the player's policy for choosing action $I(t)$ based on the information obtained so far $\{(I(s), X_{I(s)}(s), C_{I(s)}(s)) \mid s < t\}$. We evaluate policy π by using *regret*, which is defined in the following. Let $T(B) = \min\{t \mid c_t > B\}$ be the round when the player stops choosing actions. The cumulative reward of the player who follows policy π is

$$G^{\pi}(B) = \sum_{t=1}^{T(B)-1} X_{I(t)}(t).$$

Algorithm 1 Budgeted Multi-Armed Bandits [5]

```
t ← 1
c_0 ← 0
while c_{t-1} ≤ B do
    Choose an action I(t) ∈ [K].
    A cost C_{I(t)}(t) is revealed.
    c_t ← c_{t-1} + C_{I(t)}(t)
    if c_t ≤ B then
        Receive reward X_{I(t)}(t).
        t ← t + 1.
    end if
end while
```

Let $\pi^* = \arg\max_\pi \mathbb{E}[G^\pi(B)]$ be the best policy, then the performance of a policy π is measured by the *pseudo regret*

$$\overline{R}^\pi(B) = \mathbb{E}[G^{\pi^*}(B)] - \mathbb{E}[G^\pi(B)].$$

An exact evaluation of $\overline{R}^\pi(B)$ can be quite hard because in some cases choosing an action of the best reward per cost may be suboptimal.

Instead of regret for not taking the best policy π^*, we can consider regret for not choosing the best action in terms of the reward per cost. The optimal action i^* is defined as

$$i^* = \arg\max_{i\in[K]} \frac{\mu_i}{\tau_i},$$

and we will abbreviate μ_{i^*} and τ_{i^*} as μ^* and τ^* in what follows. We assume that the optimal action is unique for simplicity. Accordingly, the regret per unit cost for choosing suboptimal action i is considered to be

$$\Delta_i = \frac{\mu^*}{\tau^*} - \frac{\mu_i}{\tau_i}.$$

Let $N_i(t) = \sum_{s=1}^{t-1} \mathbb{I}\{I(s) = i\}$ denote the number of rounds action i is chosen before round t, where $\mathbb{I}\{\cdot\}$ is the indicator function. We define the following quantity

$$\tilde{R}^\pi(B) = \sum_{i:\,\Delta_i>0} \tau_i \Delta_i\, \mathbb{E}[N_i(T(B))].$$

While this quantity is not identical to the pseudo regret, the following theorem shows that it only differs by a constant from the pseudo regret[†]. Thus we again define this amount as the regret. Throughout this paper, our interest will be in this regret.

Theorem 1 (Lemma 2 by [12])**.** *For any policy π of the budgeted multi-armed bandits with budget B, we have*

$$\left|\overline{R}^\pi(B) - \tilde{R}^\pi(B)\right| \le 2\frac{\mu^*}{\tau^*}. \tag{1}$$

The regret can be minimized by a policy that minimizes $\mathbb{E}[N_i(t)]$ for suboptimal actions $i \neq i^*$ and $t \ge 1$. In the analysis, we first bound $\mathbb{E}[N_i(t)]$, then convert it into a regret bound.

[†]Note that, Lemma 2 by Xia et al. [12] shows only one side (i.e., regret upper bound) of the absolute value. However, their lemma essentially depends on the fact that the total budget used by any policy lies in $(B-1, B]$, and it is easy to use this fact to obtain a regret lower bound.

4. Policy KL-UCB-SC

In this section, we propose the policy KL-UCB-SC (KL-UCB for Stochastic Cost multi-armed bandits). Like other UCB policies, KL-UCB-SC calculates an index $U_i(t)$ of each actions i and chooses the action that maximizes the index. Namely,

$$I(t) = \arg\max_{i\in[K]} U_i(t) \tag{2}$$

where ties are broken in an arbitrary way. In the following, we define $U_i(t)$ that takes the uncertainty of the rewards and the costs into consideration. Let

$$\hat{\mu}_{i,n} = \frac{1}{n} \sum_{\substack{s:\, N_i(s)<n \\ I(s)=i}} X_i(s)$$

and

$$\hat{\tau}_{i,n} = \frac{1}{n} \sum_{\substack{s:\, N_i(s)<n \\ I(s)=i}} C_i(s).$$

Furthermore, we define the confidence region $\Phi(x, y, \delta) \subset [0,1]^2$ around point (x, y) as

$$\Phi(x, y, \delta) = \{(\mu, \tau) \mid d_{\mathrm{KL}}(x, \mu) + d_{\mathrm{KL}}(y, \tau) \le \delta\},$$

where

$$d_{\mathrm{KL}}(p, q) = p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right)$$

is the Kullback-Leibler divergence (KL divergence) of the Bernoulli distribution of mean q from the Bernoulli distribution of mean p[††]. Let $\phi(x, y, \delta)$ denote the maximum value of μ/τ among points $(\mu, \tau) \in \Phi(x, y, \delta)$:

$$\phi(x, y, \delta) = \max\{\mu/\tau \mid (\mu, \tau) \in \Phi(x, y, \delta)\},$$

Intuitively, $\phi(x, y, \delta)$ represents the point (μ, τ) of maximum μ/τ in the two dimensional region where (μ, τ) is within δ from (x, y) in terms of the pseudo-distance $d((x, y), (\mu, \tau)) = d_{\mathrm{KL}}(x, \mu) + d_{\mathrm{KL}}(y, \tau)$. The index $U_i(t)$ of an action i at round t is defined as

$$U_i(t) = \begin{cases} \phi\left(\hat{\mu}_{i,N_i(t)}, \hat{\tau}_{i,N_i(t)}, \dfrac{\ln(t)}{N_i(t)}\right) & (N_i(t) > 0) \\ \infty & (\text{otherwise}). \end{cases}$$

The advantages of the index $U_i(t)$ are that (i) the confidence bound derived from the KL divergence using Cramer-Chernoff inequality (Lemma 4) is tighter than the one derived from quadratic divergence using Hoeffding's inequality

[††]We define $0\ln(0/q) = \lim_{\epsilon\to+0} \epsilon\ln(\epsilon/q) = 0$ for any q and $p\ln(p/0) = \lim_{\epsilon\to+0} p\ln(p/\epsilon) = \infty$ for any $p > 0$.

(Fact 3), as is understood from Pinsker's inequality (Fact 2). Moreover, (ii) $U_i(t)$ incorporates the uncertainties of the reward and the cost into a single formula: the probability that both the empirical reward and the empirical cost deviate from the true mean reward and cost is intuitively small, and thus $U_i(t)$ is defined by using the sum of the two divergences.

Note that the calculation of $U_i(t)$ is a convex optimization of two variables, which can be performed by modern optimization solvers such as CVXOPT (http://cvxopt.org/).

5. Analytical Results

In this section, we analyze the KL-UCB-SC policy. Let

$$D_i = \min_{a>0} \left(d_{\mathrm{KL}}(\mu_i, a\mu^*) + d_{\mathrm{KL}}(\tau_i, a\tau^*) \right).$$

This quantity characterizes how easy it is to identify action i from the optimal one.

In the context of standard multi-armed bandit problem, Lai and Robbins [16] showed that a strongly consistent algorithm must choose action i until the number of draws $N_i(t)$ satisfies $\ln t \simeq N_i(t) d_{\mathrm{KL}}(\hat{\mu}_i, \mu^*)$. Note that the quantity $N_i(t) d_{\mathrm{KL}}(\hat{\mu}_i, \mu^*)$ can be considered as the negative logarithmic likelihood that the parameter of arm i is μ^*, and thus the strong consistency requires that the parameter of action i is not as good as the one of the optimal with significance level $1/t$. In our case, each action i has two parameters μ_i, τ_i, and D_i measures the difference between action i and an action with its expected reward $a\mu^*$ and cost $a\tau^*$. Since $a\mu^*/a\tau^* = \mu^*/\tau^*$, if action i had parameters $a\mu^*, a\tau^*$, it was as good as the optimal. Choosing action i until $\ln t \simeq N_i(t) \left(\min_{a>0} \left(d_{\mathrm{KL}}(\hat{\mu}_i, a\mu^*) + d_{\mathrm{KL}}(\hat{\tau}_i, a\tau^*) \right)\right)$ checks that the true parameters of arm i is unlikely to be $a\mu^*, a\tau^*$ for any $a > 0$, and thus its expected reward per cost does not exceed μ^*/τ^*. In this sense, D_i is the essential quantity for the budgeted bandit problem.

In Sect. 5.1, we derive an upper bound on the expected number $\mathbb{E}[N_i(T)]$ of choices of suboptimal action i. In Sect. 5.2, we introduce the notion of a strongly consistent policy and derive a lower bound of $\mathbb{E}[N_i(T)]$ for Bernoulli rewards and costs for any strongly consistent policy. On the basis of these two bounds, in Sect. 5.3, we show that (i) the regret of KL-UCB-SC for Bernoulli rewards and costs is asymptotically optimal, and (ii) its regret for any bounded rewards and costs is upper bounded by the optimal regret bound for the Bernoulli ones with the same expected rewards and expected costs.

To overview these results briefly, the technical part of the proofs are left to the appendices.

5.1 Upper Bound on the Number of Choices of Suboptimal Actions

In this section, we prove an upper bound on the expected number $\mathbb{E}[N_i(T)]$ of selections of action i in the first $T-1$ rounds for any $i \neq i^*$ and $T \geq 1$. When suboptimal action i is chosen, either of the following two events occurs: underestimation of the optimal action or overestimation of the suboptimal actions i. Lemma 1 and 2 bound these two events respectively.

The first bound is about the former event. For sufficiently small $\epsilon > 0$, let $\mu^*(\epsilon) = \mu^* - \epsilon$ and $\tau^*(\epsilon) = \tau^* + \epsilon$. Because the optimal action is expected to be chosen frequently, it is unlikely that the index of the optimal action is below $\mu^*(\epsilon)/\tau^*(\epsilon) (< \mu^*/\tau^*)$.

Lemma 1. *Let $\epsilon > 0$ be sufficiently small. Then, the following equality holds:*

$$\sum_{t=K+1}^{T} \Pr\left\{ U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \right\} = O(\epsilon^{-6}). \tag{3}$$

Proof. See Appendix B. □

The second bound is on the overestimation of the suboptimal action $i \neq i^*$.

Lemma 2. *Let $\epsilon > 0$ be sufficiently small. Then, the following equality holds for any suboptimal action $i \neq i^*$:*

$$\sum_{t=K+1}^{T} \Pr\left\{ I(t) = i, U_{i^*}(t) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \right\} = \frac{1 + O(\epsilon)}{D_i} \ln(T) + O(\epsilon^{-2}). \tag{4}$$

Proof. See Appendix C. □

By using these lemmas, we upper-bound the number of the selection of suboptimal action i.

Theorem 2. *Let the policy π be KL-UCB-SC. For any suboptimal action $i \neq i^*$ and $T \geq 1$, sufficiently small ϵ,*

$$\mathbb{E}[N_i(T)] = \frac{(1 + O(\epsilon)) \ln(T)}{D_i} + O(\epsilon^{-6})$$

holds.

Proof. $N_i(T)$ can be decomposed into the following terms:

$$N_i(T) \leq 1 + \sum_{t=K+1}^{T-1} \mathbb{I}\left\{ U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \right\} + \sum_{t=K+1}^{T-1} \mathbb{I}\left\{ I(t) = i, U_{i^*}(t) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \right\}. \tag{5}$$

Then, the proof is immediately completed by taking the expectation and using Lemma 1 (for bounding the second term) and Lemma 2 (for bounding the third term). □

Although this theorem is very close to the regret bound that we want to derive, it does not immediately derive the upper bound of expected regret as a function of budget B because in the budgeted bandit problem the number of round $T(B)$ varies among runs. In Sect. 5.3, we will revisit this theorem to derive the regret bound.

5.2 Lower Bound on the Number of Choices of Suboptimal Actions

Following a similar discussion to the one in [16], we first introduce the notion of strong consistency. Then, we derive a lower bound for the number of choices that all strongly consistent policies must have.

Definition 1 (Strong consistency)**.** *A policy is strongly consistent if, for any suboptimal action i, $a > 0$ and any distributions,*

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{T^a} \to 0.$$

Theorem 3. *Let the rewards and costs be drawn from Bernoulli distributions with parameters $\{\mu_i, \tau_i\}_{i\in[K]}$. For any strongly consistent policy, any suboptimal action i, and any $\epsilon > 0$, $\mathbb{E}[N_i(T)]$ is lower-bounded as:*

$$\mathbb{E}[N_i(T)] = (1 - o(1))\frac{\ln(T)}{D_i}.$$

Proof. See Appendix D. □

Similar to the upper bound, this lower bound on the number of choices does not immediately derive a regret lower bound as a function of the budget B. In Sect. 5.3, we convert these bounds on $\mathbb{E}[N_i(T)]$ into regret bounds.

5.3 Regret Bounds

It is not still straightforward to bound the regret by using bounds of $\mathbb{E}[N_i(T)]$ for a given T because $T(B)$ varies among runs. The following lemma, which implicitly bounds $T(B)$, relates the regret and $\mathbb{E}[N_i(T)]$ for some fixed T.

Lemma 3. *For any policy π, we have*

$$\tilde{R}^\pi(B) \le \sum_{i:\,\Delta_i>0} \tau_i\Delta_i \left(\mathbb{E}\left[N_i\left(\left\lfloor\frac{2KB}{\tau_{\min}}\right\rfloor\right)\right] + \frac{K^2}{2\tau_{\min}^2}\right) + o(1) \quad (6)$$

and

$$\tilde{R}^\pi(B) \ge \sum_{i:\,\Delta_i>0} \tau_i\Delta_i\, \mathbb{E}\left[N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor + 1\right)\right] - o(1), \quad (7)$$

where $o(1)$s are considered as functions of B.

Proof. See Appendix E. □

Lemma 3 enables us to convert the results of Theorem 2 and Theorem 3 into statements on the regret as functions of the budget B. Compared with the analysis by Xia et al. [12], that always takes the budget into consideration, Lemma 3 makes the analysis much easier.

Theorem 4. *Let π be the KL-UCB-SC. Then the regret is upper-bounded as:*

$$\tilde{R}^\pi(B) = \sum_{i:\,\Delta_i>0} \frac{\tau_i\Delta_i \ln(B)}{D_i} + o(\ln(B)).$$

Proof. The following holds by Theorem 2 with $\epsilon = (\ln(B))^{-1/7}$ for sufficiently large B and Lemma 3.

$$\begin{aligned}
&\tilde{R}^\pi(B)\\
&\le \sum_{i:\,\Delta_i>0} \tau_i\Delta_i \left((1+O(\epsilon))\frac{\ln\left(\frac{2KB}{\tau_{\min}}\right)}{D_i} + O(\epsilon^{-6}) + \frac{K^2}{2\tau_{\min}^2}\right) + o(1)\\
&= \sum_{i:\,\Delta_i>0} \tau_i\Delta_i(1+O(\epsilon))\frac{\ln(B) + \ln\left(\frac{2K}{\tau_{\min}}\right)}{D_i} + O(\epsilon^{-6})\\
&\quad \left(\text{by } O(\epsilon^{-6}) + \frac{K^2}{2\tau_{\min}^2}\sum_{i:\,\Delta_i>0}\tau_i\Delta_i + o(1) = O(\epsilon^{-6})\right)\\
&= \sum_{i:\,\Delta_i>0} \frac{\tau_i\Delta_i\ln(B)}{D_i} + o(\ln(B)).\\
&\quad \left(\text{by } O(\epsilon)\ln(B) = O(\epsilon^{-6}) = O\left((\ln(B))^{6/7}\right) = o(\ln(B))\right)
\end{aligned}$$

□

Theorem 5. *Let the rewards and costs be drawn from Bernoulli distributions with parameters $\{\mu_i, \tau_i\}_{i\in[K]}$. For any strongly consistent policy π, the regret is lower-bounded as:*

$$\tilde{R}^\pi(B) = \sum_{i:\,\Delta_i>0} \frac{\tau_i\Delta_i \ln(B)}{D_i} - o(\ln(B)).$$

Proof. Similar to the proof of Theorem 4, using Theorem 3 and Lemma 3, we find that

$$\begin{aligned}
\tilde{R}^\pi(B) &\ge \sum_{i:\,\Delta_i>0} \frac{(1-o(1))\tau_i\Delta_i}{D_i}\ln\left(\frac{B}{2K\tau_{\max}}\right) - o(1)\\
&= \sum_{i:\,\Delta_i>0} \frac{(1-o(1))\tau_i\Delta_i}{D_i}(\ln(B) - \ln(2K\tau_{\max})) - o(1)\\
&= \sum_{i:\,\Delta_i>0} \frac{\tau_i\Delta_i\ln(B)}{D_i} - o(\ln(B))
\end{aligned}$$

□

These two theorems immediately give us the following optimality of the KL-UCB-SC:

Corollary 1. *On budgeted bandits with stochastic costs, if all reward/cost distributions are Bernoulli, KL-UCB-SC is asymptotically optimal. That is, the coefficient of the logarithmic factor of regret as a function of budget B cannot be improved.*

5.4 Comparison with Other Policies

For a budgeted multi-armed bandits with stochastic costs, a number of policies, such as UCB-BV1 [5] and BTS [12] have been proposed. The same as we did, the upper bounds of expected regret on their policies have been proved.

Remark 1. *For the budgeted multi-armed bandits, the leading terms in the known upper bound of expected regret are:*

BTS [12]

$$\sum_{i:\,\Delta_i>0} 2\left(\frac{\mu^*}{\tau^*}+1\right)^2 \frac{\ln(B)}{\tau_i\Delta_i}.$$

UCB-BV1 [5] ($\lambda = \tau_{\min}$)

$$\sum_{i:\,\Delta_i>0} \left(\frac{2+2/\tau_{\min}+\Delta_i}{\tau_{\min}}\right)^2 \frac{\tau_i \ln(B)}{\Delta_i},$$

derived from their intermediate result

$$\mathbb{E}[N_i(T(B))] = \left(\frac{2+2/\tau_{\min}+\Delta_i}{\Delta_i\tau_{\min}}\right)^2 \ln(T(B)) + O(1)$$

and Eq. (6).

These bounds and the bound we derived for KL-UCB-SC are not naturally comparable. In order to make them comparable, we give *the upper bound of the upper bound* written in terms of Δ_i.

Lemma 4.

$$D_i \geq \frac{2(\tau^*)^2}{(\mu^*)^2+(\tau^*)^2}\tau_i^2\Delta_i^2.$$

Proof. D_i can be expressed as

$$\begin{aligned} D_i &= \min_a\{d_{\mathrm{KL}}(\mu_i, a\mu^*) + d_{\mathrm{KL}}(\tau_i, a\tau^*)\} \\ &\geq \min_a\{2(a\mu^*-\mu_i)^2 + 2(a\tau^*-\tau_i)^2\} \\ &= \frac{2(\mu_i\tau^*-\mu^*\tau_i)^2}{(\mu^*)^2+(\tau^*)^2} = \frac{2(\Delta_i\tau_i\tau^*)^2}{(\mu^*)^2+(\tau^*)^2}, \end{aligned}$$

where the inequality holds by Pinsker's inequality and the right hand side of the second last equality is the square of the distance from a point (μ_i, τ_i) to a line $x = (\mu^*/\tau^*)y$ in the x-y plane. □

This lemma makes clear the relation of our result with the previously known ones: the upper bound of the regret for KL-UCB-SC proved in Theorem 4 can be loosened to the one in the following corollary.

Corollary 2. *Let π be the KL-UCB-SC. Then the regret is upper-bounded as:*

$$\tilde{R}^\pi(B) = \sum_{i:\,\Delta_i>0} \frac{1}{2}\left(\left(\frac{\mu^*}{\tau^*}\right)^2+1\right)\frac{\ln(B)}{\tau_i\Delta_i} + o(\ln(B)) \quad (8)$$

for any bounded distribution.

Remark 2. *By comparing bound* (8) *with those in Remark 1, the regret bound improves at least by a factor of four on the best known one (proved for BTS [12]). Although this fact does not conclude that these previously-proposed policies are not asymptotically optimal even if the reward/cost distributions are Bernoulli (in other words, they have worse regret than KL-UCB-SC), it implies that our analysis is essentially tighter than the existing ones.*

6. Experiments

6.1 Policies

We compared the following six policies in our experiments: KL-UCB-SC, KL-UCB-SC+, BTS [12], PD-BwK [8], UCB-BV1 [5] and EF-KUBE. KL-UCB-SC+ is a version of KL-UCB-SC that uses $\ln(t/N_i(t))$ instead of $\ln(t)$, which is inspired by the KL-UCB+ policy for the classical multi-armed bandits [18]. The introduction of the denominator $N_i(t)$ does not change the asymptotic property because for any suboptimal action i, $N_i(t) = O(\ln t)$ holds almost surely and thus $\ln(t/N_i(t))$ is still $O(\ln t)$, while it generally improves a finite-time performance [9], [18]. BTS is an extension of Thompson sampling that draws reward and cost parameters from Beta distributions at each round and greedily chooses the estimated optimal action based on the drawn parameters. Note that BTS corresponds to the posterior sampling if and only if rewards and costs are Bernoulli. Different from BTS, the last three policies also belong to the UCB family and choose action i with the highest index, where the indices of action i at round t in the individual policies are defined as follows:

PD-BwK

$$\frac{\left[\hat{\mu}_{i,N_i(t)} + \left(\sqrt{C_{\mathrm{rad}}\hat{\mu}_{i,N_i(t)}/N_i(t)} + C_{\mathrm{rad}}/N_i(t)\right)\right]_{\leq 1}}{\left[\hat{\tau}_{i,N_i(t)} - \left(\sqrt{C_{\mathrm{rad}}\hat{\tau}_{i,N_i(t)}/N_i(t)} + C_{\mathrm{rad}}/N_i(t)\right)\right]_{\geq 0}}$$

UCB-BV1

$$\frac{\hat{\mu}_{i,N_i(t)}}{\hat{\tau}_{i,N_i(t)}} + \frac{(1+1/\lambda)\sqrt{\ln(t)/N_i(t)}}{\left[\lambda - \sqrt{\ln(t)/N_i(t)}\right]_{\geq 0}}$$

EF-KUBE

$$\frac{1}{\hat{\tau}_{i,N_i(t)}}\left(\hat{\mu}_{i,N_i(t)} + \sqrt{\frac{\ln(t)}{2N_i(t)}}\right)$$

where $[x]_{\geq\theta} = \max\{x,\theta\}$, $[x]_{\leq\theta} = \min\{x,\theta\}$. The parameters of PD-BwK and UCB-BV1 are set to $C_{\mathrm{rad}} = 0.25\ln(KB)$ and $\lambda = \tau_{\min}$ respectively[†]. Note that EF-KUBE (Empirical Fractional KUBE) is a modification of

[†]To guarantee the regret upper bound shown in [8], setting $C_{\mathrm{rad}} = \Theta(\ln(KB))$ is necessary for PD-BwK and $C_{\mathrm{rad}} = 0.25\ln(KB)$ is the setting used in the comparison experiment in [12]. Setting $\lambda = \tau_{\min}$ is optimal for UCB-BV1 in terms of the regret upper bound shown in [5].

Table 1 Reward and cost distributions of five actions.

(a) Scenario M: Medium mean reward and cost.

i	$X_i(t)$	$C_i(t)$	$\frac{\mu_i}{\tau_i}$
1	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	1
2	$\mathcal{B}(0.4)$	$\mathcal{B}(0.4)$	1
3	$\mathcal{B}(0.6)$	$\mathcal{B}(0.6)$	1
4	$\mathcal{B}(0.4)$	$\mathcal{B}(0.6)$	$0.\dot{6}$
5	$\mathcal{B}(0.6)$	$\mathcal{B}(0.4)$	1.5

(b) Scenario LR: Low mean reward.

i	$X_i(t)$	$C_i(t)$	$\frac{\mu_i}{\tau_i}$
1	$\mathcal{B}(0.02)$	$\mathcal{B}(0.4)$	0.05
2	$\mathcal{B}(0.03)$	$\mathcal{B}(0.4)$	0.075
3	$\mathcal{B}(0.04)$	$\mathcal{B}(0.8)$	0.05
4	$\mathcal{B}(0.06)$	$\mathcal{B}(0.8)$	0.075
5	$\mathcal{B}(0.08)$	$\mathcal{B}(0.8)$	0.1

(c) Scenario H: High mean reward and cost.

i	$X_i(t)$	$C_i(t)$	$\frac{\mu_i}{\tau_i}$
1	$\mathcal{B}(0.9)$	$\mathcal{B}(0.9)$	1.0
2	$\mathcal{B}(0.9)$	$\mathcal{B}(0.9)$	1.0
3	$\mathcal{B}(0.9)$	$\mathcal{B}(0.9)$	1.0
4	$\mathcal{B}(0.9)$	$\mathcal{B}(0.9)$	1.0
5	$\mathcal{B}(0.9)$	$\mathcal{B}(0.8)$	1.125

(d) Scenario FR: Fixed reward.

i	$X_i(t)$	$C_i(t)$	$\frac{\mu_i}{\tau_i}$
1	$\mathcal{B}(1)$	$\mathcal{B}(0.4)$	2.5
2	$\mathcal{B}(1)$	$\mathcal{B}(0.4)$	2.5
3	$\mathcal{B}(1)$	$\mathcal{B}(0.4)$	2.5
4	$\mathcal{B}(1)$	$\mathcal{B}(0.4)$	2.5
5	$\mathcal{B}(1)$	$\mathcal{B}(0.3)$	$3.\dot{3}$

fractional KUBE [4]; the original fractional KUBE uses time-invariant costs τ_i instead of $\hat{\tau}_{i,N_i(t)}$.

6.2 Synthetic Dataset

First we simulate a 5-armed bandit game for four scenarios. Budget B is set to $10000 = 10^4$. We use Bernoulli distributions $\mathcal{B}(\mu)$ of mean μ. The rewards and costs of the five actions are generated according to the distributions shown in Table 1. The reward and cost means of actions in Scenario M are selected from the three values $0.4, 0.5, 0.6$ around the center of the range $[0, 1]$. The reward means of Scenario LR are low for all the actions, which is typical when the number of clicks is used as the cumulative reward in the applications to recommendation and advertising. In this case, estimation of reward means are difficult because non-zero rewards can be rarely obtained in Bernoulli distribution. Scenario H and FR are special cases. All the reward and cost means are very high in Scenario H, and all the rewards are fixed while all the cost means are medium in Scenario FR. In the both scenario, reward and cost means are the same for all the actions except cost mean for action 5. The optimal action i^* is 5 in all scenarios.

Table 2 shows the empirical regret averaged over 1000 runs for $B = 10^4$. The empirical regret $\hat{R}(B)$ is calculated as

$$\hat{R}(B) = \sum_{i\,:\,\Delta_i > 0} \tau_i \Delta_i N_i(T(B)).$$

Figure 1 shows the increase of the empirical regret (y-axis) against the budget (x-axis) for the six policies.

KL-UCB-SC+ and BTS stably perform well. They perform third best at worst for the four scenarios. (See Table 2.) Compared with BTS, KL-UCB-SC+ performs slightly better in Scenario LR and H, slightly worse in Scenario M, and as well in Scenario FR.

Note that Thompson sampling and KL-UCB for the classical bandit problem are both asymptotically optimal. Although BTS has not been analyzed in depth yet, taking the fact that BTS performs very close to KL-UCB-SC+, it is possible that BTS is also asymptotically optimal in the budgeted bandits.

Table 2 Averaged empirical regrets for $B = 10^4$ (The tiny parenthesized numbers are ranks in ascending order.)

Scenario	M	LR	H	FR
KL-UCB-SC	182.82 (5)	89.63 (4)	105.47 (3)	469.92 (5)
KL-UCB-SC+	116.48 (3)	50.73 (2)	67.93 (1)	283.29 (3)
BTS	112.21 (2)	55.83 (3)	72.34 (2)	281.53 (2)
PD-BwK	152.02 (4)	48.36 (1)	263.02 (5)	200.84 (1)
UCB-BV1	2498.65 (6)	268.21 (6)	691.57 (6)	3230.75 (6)
EF-KUBE	82.43 (1)	180.31 (5)	127.32 (4)	360.21 (4)

Table 3 Average computation time per decision of each policy on scenario LR.

Policy	Time (μs)
KL-UCB-SC	19 741.1
KL-UCB-SC+	21 473.8
BTS	14.5
PD-BwK	4.2
UCB-BV1	2.3
EF-KUBE	2.6

PD-BwK performs best in Scenario LR and FR but poorly performs in Scenario H. PD-BwK has parameter $C_{\mathtt{rad}}$ on which its performance depend. Parameter $C_{\mathtt{rad}} = 0.25 \ln KB$, which is used in our experiments, are considered to be good for Scenario LR and FR, and bad for Scenario H. Parameter tuning may improve its performance for Scenario H, but such a troublesome task is one of the demerits of PD-BwK.

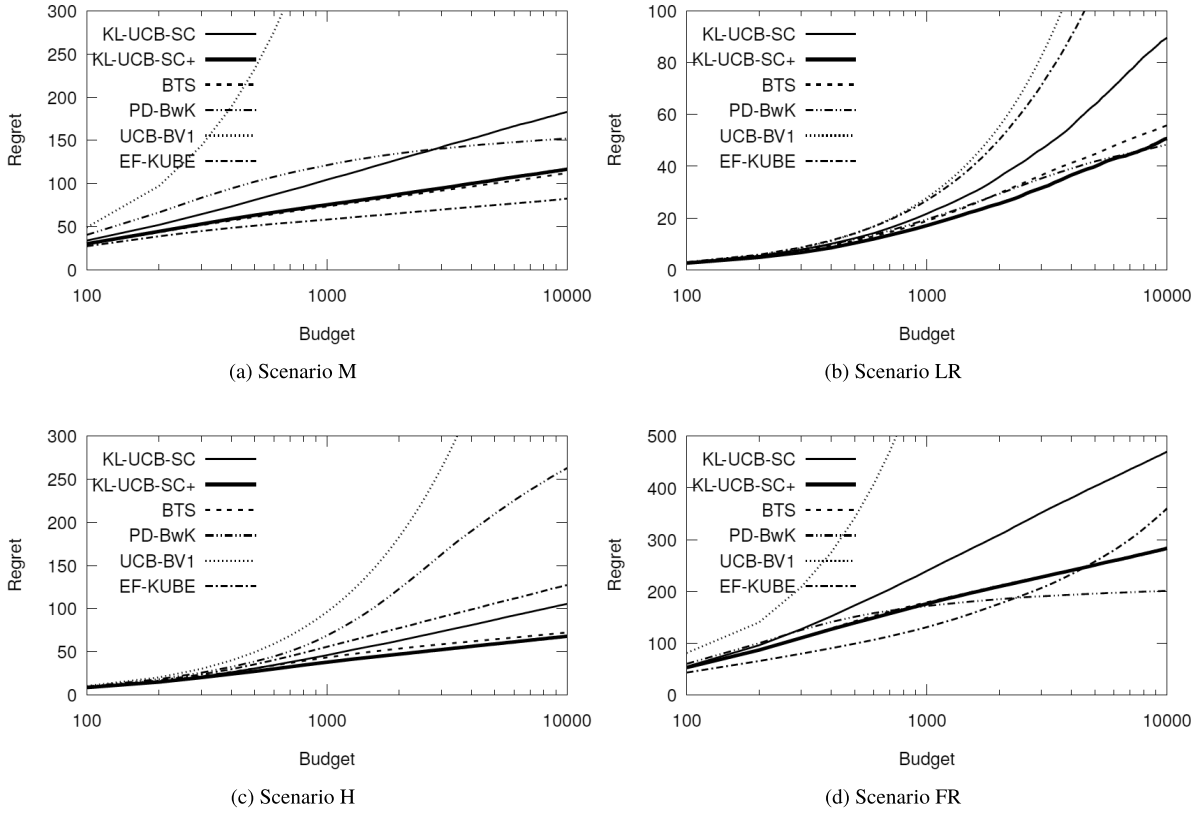
EF-KUBE performs best in Scenario M but its performance is not good for the other scenarios. In EF-KUBE, its index does not consider how much confidence the estimated cost mean has, and uses the estimated cost mean directly. As a result, in the case of over-estimation of the cost mean, it takes many rounds to recover its estimation. In Scenario FR, the regret increasing rate of EF-KUBE is high, which is caused by a small fraction of runs in which over-estimation of the cost mean of the best action occurs.

UCB-BV1 performed worst for all the scenarios.

In addition, the average computation time per decision of each policy on scenario LR is shown in Table 3. By our current implemantetion of KL-UCB-SC(+) using convex optimization, it takes about 0.004 s to calculate the index of one action, which is, unfortunately, more than 1000 times slower compared to other policies.

6.3 Real-World Dataset

In order to show effectiveness for real-world applications, we also conducted simulations of real-time bidding advertising based on real-world dataset provided by iPinYou [19]. In real-time bidding advertising, every time when a user visits a website containing an ad space, an auction for the ad space is held in real time. In the auction, advertisers bid their prices by the bidding rules (policies) that they set in advance, and one of them wins and pays the second bidding price instead



(a) Scenario M

(b) Scenario LR

(c) Scenario H

(d) Scenario FR

Fig. 1 Increase of empirical regret averaged over 1000 independent runs. A base-10 log scale is used for the x-axis. In Scenario M and FR, the curve for BTS is overwritten by the curve for KL-UCB-SC+ because their differences are within ±6.0 and ±4.0, respectively, for all the budgets.

of his/her bidding price for the impression of his/her ad on the site at that time. The second bidding price cannot be controlled by the winner, so the impression cost can be seen as a random variable. In our experimental setting, we assume that every user belongs to one of K categories and the visiting user's category is revealed to the advertisers in the auction. The player in our simulation is one of the advertisers who wins the auctions for visits of users of a target category only. The player is assumed to be able to win any auction by bidding a price enough to win. So, the player repeats to decide a target user category, bid at the auction for the next visit of a user of the target category, win the auction, pay the cost of the impression, and receive an information whether the displayed ad is clicked or not. In this setting, we simulated a budgeted bandit problem in which a player tries to maximize the number of clicks given a fixed amount of budget by targeting users via their categories.

As the distributions of reward $X_i(t)$ and cost $C_i(t)$ for user category i and the player's tth win of the auctions, we used Bernoulli distribution $\mathcal{B}(\mu_i)$ and lognormal distribution[†] $\Lambda(\nu_i, \sigma_i^2)$, respectively, where $X_i(t) = 1$ or 0 means that the player's ad is clicked or not by the user of category i for the tth impression of the player's ad, and $C_i(t)$ is the tth impression cost.

[†]We used lognormal distribution as a cost distribution following a literature by Ostrovsky and Schwartz [20].

We set the number of user categories to 20 and all the distribution parameters μ_i, ν_i, σ_i^2 $(i = 1, \ldots, 20)$ are estimated from the training dataset of the second season of iPin You global RTB bidding algorithm competition. Using the impression log, we selected the top-twenty most frequently appeared sets of user profile ids[††] as actions (user categories). Reward mean μ_i of each action i was set to the empirical click-through-rate (i.e, probability of click) of the corresponding user category (the set of user profile ids), which was calculated from the impression and click logs. The click-through-rate of actions are in 1–6.2 percent[†††]. The cost distribution parameters ν_i and σ_i^2 of each action i are also set to the empirical mean and variance of the logarithmic paying price of the impression for the corresponding user category, where paying prices are provided by the impression log. The mean costs of the actions are in 60–102. (See Table A· 1 for the parameters that we used in our simulations.)

Given this, we set $B = 1000000 = 10^6$ in this dataset. The policies compared are the same as the ones of the pre-

[††]Several profile ids are assigned to one user according to his/her demographic and geographic information, long-term interest and in-market purchase perspectives.

[†††]Because the click-through-rates of display advertisements are usually very small, we multiplied 100 to each of them, and thus each selection of an action approximately simulates a hundred impressions.

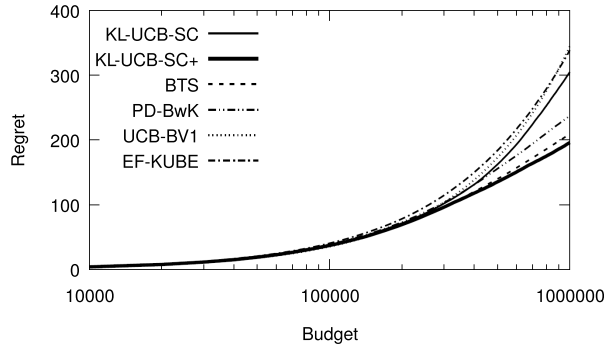


Fig. 2 Increase of empirical regret on real-world dataset averaged over 100 independent runs.

vious simulations.

The result averaged over 100 runs is summarized in Fig. 2. In this experiment, KL-UCB-SC+ also performed the best which is followed by BTS.

7. Conclusion and Future Work

We proposed a policy named KL-UCB-SC for budgeted multi-armed bandits with stochastic action costs. We derived a regret upper bound for it that holds for any bounded rewards and costs. The regret bound is optimal in the case that the rewards and costs are drawn from Bernoulli distributions, which is, to the best of our knowledge, the first result that addresses optimality in budgeted multi-armed bandits. Furthermore, we demonstrated that the performance of KL-UCB-SC+, which is a variant of KL-UCB-SC, is comparable to that of the state-of-the-art policies in numerical experiments.

In order to calculate the index of KL-UCB-SC, we had to solve a convex optimization problem, which takes more time than existing policies. An efficient way to compute the index would be an interesting future work.

References

[1] W.R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," Biometrika, vol.25, no.3-4, pp.285–294, 1933.
[2] H. Robbins, "Some aspects of the sequential design of experiments," Bull. Am. Math. Soc., vol.58, no.5, pp.527–535, 1952.
[3] L. Tran-Thanh, A. Chapman, E.M. de Cote, A. Rogers, and N.R. Jennings, "ϵ-first policies for budget-limited multi-armed bandits," Proc. 24th AAAI Conference on Artificial Intelligence, pp.1211–1216, 2010.
[4] L. Tran-Thanh, A. Rogers, A. Chapman, and N.R. Jennings, "Knapsack based optimal policies for budget-limited multi-armed bandits," Proc. 26th AAAI Conference on Artifical Intelligence, pp.1134–1140, 2012.
[5] W. Ding, T. Qin, X.D. Zhang, and T.Y. Liu, "Multi-armed bandit with budget constraint and variable costs," Proc. 27th AAAI Conference on Artificial Intelligence, pp.232–238, 2013.
[6] K. Amin, M. Kearns, P. Key, and A. Schwaighofer, "Budget optimization for sponsored search: Censored learning in mdps," Proc. 28th Conference on Uncertainty in Artificial Intelligence, pp.54–63, 2012.
[7] L. Tran-Thanh, L. Stavrogiannis, V. Naroditskiy, V. Robu, N. R Jennings, and P. Key, "Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions," Proc. 30th Conference on Uncertainty in Artificial Intelligence, pp.809–818, 2014.
[8] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, "Bandits with knapsacks," Proc. 54th IEEE Annual Symposium on Foundation of Computer Science, pp.207–216, 2013.
[9] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," Proc. 22nd International Conference on Algorithmic Learning Theory, pp.359–376, 2011.
[10] O. Cappé, A. Garivier, O.A. Malliard, R. Munos, and G. Stoltz, "Kullback-leibler upper confidence bounds for optimal sequential allocation," Ann. Statist., vol.41, no.3, pp.1516–1541, 2013.
[11] T.L. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," Ann. Statist., vol.15, no.3, pp.1091–1114, 1987.
[12] Y. Xia, H. Li, T. Qin, N. Yu, and T.Y. Liu, "Thompson sampling for budgeted multi-armed bandits," Proc. 24th International Joint Conference on Artificial Intelligence, pp.3960–3966, 2015.
[13] P. Auer, N. Cesa-Bianchi, Y. Freund, and E. Schapire, Robert, "The nonstochastic multiarmed bandit problem," SIAM J. Comput., vol.32, no.1, pp.48–77, 2003.
[14] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," Mach. Learn., vol.47, no.2-3, pp.235–256, 2002.
[15] J. Honda and A. Takemura, "An asymptotically optimal bandit algorithm for bounded support models," Proc. 23rd Annual Conference on Learning Theory, pp.67–79, 2010.
[16] T.L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," Adv. Appl. Math., vol.6, no.1, pp.4–22, 1985.
[17] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," Proc. 23rd International Conference on Algorithmic Learning Theory, pp.199–213, 2012.
[18] E. Kaufmann, Analyse de stratégies bayésiennes et fréqentistes pour l'allocation séquentille de ressources, Ph.D. thesis, Telecom ParisTech, 2014.
[19] W. Zhang, S. Yuan, J. Wang, and X. Shen, "Real-time bidding benchmarking with ipinyou dataset," arXiv:1407.7073, 2014.
[20] M. Ostrovsky and M. Schwartz, "Reserve prices in internet advertising auctions: A field experiment," Proc. 12th ACM Conference on Electronic Commerce, pp.59–60, 2011.
[21] W. Hoeffding, "Probability inequalities for sums of bounded random variables," J. Am. Stat. Assoc., vol.58, no.301, pp.13–30, 1963.
[22] A. Dembo and O. Zeitouni, Large Deviations Techniques and Applications, Springer-Verlag Berlin Heidelberg, 2010.
[23] S. Bubeck, Bandit Games and Clustering Foundations, Ph.D. thesis, Université Lille 1, 2010.

Appendix A: Important Inequalities

The following four inequalities are important inequalities to prove our bounds.

Fact 1 (Markov's inequality)**.** *For any random variable X with a non-negative support and $a > 0$,*

$$\mathbb{E}[X] \geq a \cdot \Pr\{X \geq a\}$$

holds.

Fact 2 (Pinsker's inequality)**.** *For all $p, q \in [0, 1]$, the following inequality holds:*

$$d_{\mathrm{KL}}(p, q) \geq 2(p - q)^2.$$

Fact 3 (Hoeffding's inequality [21])**.** *Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$,*

where $X_1, \ldots, X_n$ are i.i.d. random variables taking values in $[0, 1]$. Let μ be the their common mean. Then, for any $a \geq 0$,

$$\Pr\{\hat{\mu}_n \geq \mu + a\} \leq \mathrm{e}^{-2na^2}$$

and

$$\Pr\{\hat{\mu}_n \leq \mu - a\} \leq \mathrm{e}^{-2na^2}.$$

Fact 4 (Cramer-Chernoff inequality for bounded random variables.)**.** *Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are i.i.d. random variables taking values in $[0, 1]$. Let μ be the their common mean. Then, for any $x > \mu$,*

$$\Pr\{\hat{\mu}_n \geq \mu + a\} \leq \mathrm{e}^{-nd_{\mathrm{KL}}(\mu+a,\mu)}, \tag{A·1}$$

and for any $x < \mu$,

$$\Pr\{\hat{\mu}_n \leq \mu - a\} \leq \mathrm{e}^{-nd_{\mathrm{KL}}(\mu-a,\mu)}. \tag{A·2}$$

Proof. By Eq. (2.2.12) and (2.2.13) in [22], for any $x > \mu$,

$$\Pr\{\hat{\mu}_n \geq x\} \leq \mathrm{e}^{-n\Lambda^*(x)},$$

and for any $x < \mu$,

$$\Pr\{\hat{\mu}_n \leq x\} \leq \mathrm{e}^{-n\Lambda^*(x)},$$

where $\Lambda(\lambda) = \ln \mathbb{E}[\mathrm{e}^{\lambda X_i}]$ and

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\}.$$

By Lemma 9 in [9], $\mathbb{E}[\mathrm{e}^{\lambda X}] \leq 1 - \mu + \mathrm{e}^{\lambda}$ holds for any $\lambda \in \mathbb{R}$, and the proof is completed by using the convexity of $\Lambda^*(\lambda)$. □

Appendix B: Proof of Lemma 1

We bound $\sum_{t=K+1}^T \Pr\left\{U_{i^*}(t) \leq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}$ by the sum of three probability summations, and further bound each of them. We first prove Lemma 5, which bounds each of the three probability summations. Then, we prove Lemma 1 using Lemma 5. The next two propositions are necessary for the proof of Lemma 5. Note that function f defined in Proposition 1 is used in the proof of Lemma 5.

Proposition 1 ([15])**.** *Let $f(\mu, \mu_1) = (\mu - \mu_1)^2/(2\mu(1-\mu_1))$. Then,*

$$d_{\mathrm{KL}}(\mu_2, \mu) - d_{\mathrm{KL}}(\mu_2, \mu_1) \geq f(\mu, \mu_1) > 0 \tag{A·3}$$

holds for any $0 < \mu_2 \leq \mu_1 < \mu < 1$ and

$$d_{\mathrm{KL}}(\tau_2, \tau) - d_{\mathrm{KL}}(\tau_2, \tau_1) \geq f(\tau_1, \tau) > 0 \tag{A·4}$$

holds for any $0 < \tau < \tau_1 \leq \tau_2 < 1$.

Proof. The proof of Eq. (A·3) is in the original paper [15]. Equation (A·4) can be proved in the same way. □

Proposition 2. *For any $a > 0$, the following inequality holds:*

$$\sum_{k=1}^{\infty} k^m \mathrm{e}^{-ak} \leq \begin{cases} \dfrac{m!\mathrm{e}^a}{a^{m+1}} & (m \in \mathbb{N}) \\ \dfrac{1}{a} & (m = 0). \end{cases}$$

Proof. Since $k^m e^{ak} \leq (x+1)^m e^{-ax}$ holds for all $k - 1 \leq x \leq k$,

$$\begin{aligned} \sum_{k=1}^{\infty} k^m \mathrm{e}^{-ak} &\leq \int_0^{\infty} (x+1)^m \mathrm{e}^{-ax} \mathrm{d}x \\ &\leq \frac{\mathrm{e}^a}{a} \int_a^{\infty} \left(\frac{z}{a}\right)^m \mathrm{e}^{-z} \mathrm{d}z \\ &\qquad \text{(by substitution } z = a(x+1)) \\ &\leq \begin{cases} \dfrac{\mathrm{e}^a \Gamma(m+1)}{a^{m+1}} & (m \in \mathbb{N}) \\ \dfrac{1}{a} & (m = 0). \end{cases} \end{aligned}$$

□

Now we prove Lemma 5.

Lemma 5. *Let $\epsilon > 0$ be sufficiently small. Let $F_n^{\mu}(x) = \Pr\{\hat{\mu}_{i^*,n} \leq x\}$ and $F_n^{\tau}(y) = \Pr\{\hat{\tau}_{i^*,n} \geq y\}$. Then, the following equations hold.*

$$\sum_{n=1}^{\infty} \int_{x=0}^{\mu^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(x, \mu^*(\epsilon)\right)\right) \mathrm{d}F_n^{\mu}(x) = O(\epsilon^{-4}),$$

$$\sum_{n=1}^{\infty} \int_{y=1}^{\tau^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) \mathrm{d}F_n^{\tau}(y) = O(\epsilon^{-4}),$$

$$\begin{aligned} &\sum_{n=1}^{\infty} \left(\int_{x=0}^{\mu^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(x, \mu^*(\epsilon)\right)\right) \mathrm{d}F_n^{\mu}(x) \right. \\ &\qquad \left. \cdot \int_{y=1}^{\tau^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) \mathrm{d}F_n^{\tau}(y) \right) \\ &= O(\epsilon^{-6}). \end{aligned}$$

Proof. We first derive the second equation. By using integration by parts, we obtain

$$\begin{aligned} &\int_{y=1}^{\tau^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) \mathrm{d}F_n^{\tau}(y) \\ &= \left[\exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) F_n^{\tau}(y)\right]_{y=1}^{\tau^*(\epsilon)} \\ &\quad + \int_{y=1}^{\tau^*(\epsilon)} -\frac{\mathrm{d}\exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right)}{\mathrm{d}y} F_n^{\tau}(y) \mathrm{d}y. \end{aligned}$$

The first term can be bounded as

$$\begin{aligned} &\left[\exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) F_n^{\tau}(y)\right]_{y=1}^{\tau^*(\epsilon)} \\ &= F_n^{\tau}(\tau * (\epsilon)) - \exp\left(nd_{\mathrm{KL}}\left(1, \tau^*(\epsilon)\right)\right) F_n^{\tau}(1) \\ &\leq F_n^{\tau}(\tau^*(\epsilon)) = \Pr\{\hat{\tau}_{i^*,n} \geq \tau^* + \epsilon\} \end{aligned}$$

$$\leq \exp(-nd_{\mathrm{KL}}(\tau^*(\epsilon), \tau^*))$$

by using Cramer-Chernoff inequality (Fact 4). The second term is bounded as follows:

$$\begin{aligned}
&\int_{y=1}^{\tau^*(\epsilon)} -\frac{\mathrm{d}\exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right)}{\mathrm{d}y} F_n^{\tau}(y)\mathrm{d}y \\
&= \int_{y=\tau^*(\epsilon)}^{1} n\left(\ln\left(\frac{y}{\tau^*(\epsilon)}\right) - \ln\left(\frac{1-y}{1-\tau^*(\epsilon)}\right)\right) \\
&\qquad \cdot \exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) F_n^{\tau}(y)\mathrm{d}y \\
&\leq \int_{y=\tau^*(\epsilon)}^{1} n\left(\ln\left(\frac{y}{\tau^*(\epsilon)}\right) - \ln\left(\frac{1-y}{1-\tau^*(\epsilon)}\right)\right) \\
&\qquad \cdot \exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right) - nd_{\mathrm{KL}}\left(y, \tau^*\right)\right) \mathrm{d}y \\
&\qquad \text{(by Cramer-Chernoff inequality (Fact 4))} \\
&\leq \int_{y=\tau^*(\epsilon)}^{1} n\left(\ln\left(\frac{y}{\tau^*(\epsilon)}\right) - \ln\left(\frac{1-y}{1-\tau^*(\epsilon)}\right)\right) \\
&\qquad \cdot \exp\left(-nf\left(\tau^*(\epsilon), \tau^*\right)\right) \mathrm{d}y \\
&\qquad \text{(by Proposition 1)} \\
&= n\ln\left(\frac{1}{\tau^*(\epsilon)}\right)\exp(-nf(\tau^*(\epsilon), \tau^*)).
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
&\sum_{n=1}^{\infty}\int_{y=1}^{\tau^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) \mathrm{d}F_n^{\tau}(y) \\
&\leq \sum_{n=1}^{\infty}\left(\exp(-nd_{\mathrm{KL}}(\tau^*(\epsilon), \tau^*)) \right. \\
&\qquad \left. + \ln\left(\frac{1}{\tau^*(\epsilon)}\right) n\exp(-nf(\tau^*(\epsilon), \tau^*))\right) \\
&\leq \underbrace{\frac{1}{d_{\mathrm{KL}}(\tau^*(\epsilon), \tau^*)}}_{O(\epsilon^{-2})\text{ (by Pinsker's inequality)}} \\
&\qquad + \underbrace{\ln\left(\frac{1}{\tau^*(\epsilon)}\right)\frac{\exp(f(\tau^*(\epsilon), \tau^*))}{f^2(\tau^*(\epsilon), \tau^*)}}_{O(\epsilon^{-4})\text{ (by the definition of } f)} \\
&\qquad \text{(by Proposition 2 for } m = 0, 1) \\
&= O(\epsilon^{-4}).
\end{aligned}$$

Using the same argument yields

$$\sum_{n=1}^{\infty}\int_{x=0}^{\mu^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(x, \mu^*(\epsilon)\right)\right) \mathrm{d}F_n^{\mu}(x) = O(\epsilon^{-4}).$$

Similarly, we have

$$\begin{aligned}
&\sum_{n=1}^{\infty}\int_{x=0}^{\mu^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(x, \mu^*(\epsilon)\right)\right) \mathrm{d}F_n^{\mu}(x) \\
&\qquad \cdot \int_{y=1}^{\tau^*(\epsilon)} \exp\left(nd_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) \mathrm{d}F_n^{\tau}(y) \\
&\leq \sum_{n=1}^{\infty}\left(\exp(-nd_{\mathrm{KL}}(\mu^*(\epsilon), \mu^*)) \right. \\
&\qquad \left. + \ln\left(\frac{1}{1-\mu^*(\epsilon)}\right) n\exp(-nf(\mu^*, \mu^*(\epsilon)))\right) \\
&\qquad \cdot \left(\exp(-nd_{\mathrm{KL}}(\tau^*(\epsilon), \tau^*)) \right. \\
&\qquad \left. + \ln\left(\frac{1}{\tau^*(\epsilon)}\right) n\exp(-nf(\tau^*(\epsilon), \tau^*))\right) \\
&\leq \frac{1}{d_{\mathrm{KL}}(\mu^*(\epsilon), \mu^*) + d_{\mathrm{KL}}(\tau^*(\epsilon), \tau^*)} \\
&\qquad + \ln\left(\frac{1}{\tau^*(\epsilon)}\right)\frac{\exp\left(d_{\mathrm{KL}}(\mu^*(\epsilon), \mu^*) + f(\tau^*(\epsilon), \tau^*)\right)}{\left(d_{\mathrm{KL}}(\mu^*(\epsilon), \mu^*) + f(\tau^*(\epsilon), \tau^*)\right)^2} \\
&\qquad + \ln\left(\frac{1}{1-\mu^*(\epsilon)}\right) \\
&\qquad\qquad \cdot \frac{\exp\left(d_{\mathrm{KL}}(\tau^*(\epsilon), \tau^*) + f(\mu^*, \mu^*(\epsilon))\right)}{\left(d_{\mathrm{KL}}(\tau^*(\epsilon), \tau^*) + f(\mu^*, \mu^*(\epsilon))\right)^2} \\
&\qquad + 2\ln\left(\frac{1}{1-\mu^*(\epsilon)}\right)\ln\left(\frac{1}{\tau^*(\epsilon)}\right) \\
&\qquad\qquad \cdot \frac{\exp\left(f(\mu^*, \mu^*(\epsilon)) + f(\tau^*(\epsilon), \tau^*)\right)}{\left(f(\mu^*, \mu^*(\epsilon) + f(\tau^*(\epsilon), \tau^*)\right)^3} \\
&\qquad \text{(by Proposition 2 for } m = 0, 1 \text{ and } 2) \\
&= O(\epsilon^{-6}). \\
&\qquad \text{(by Pinsker's inequality and the definition of } f)
\end{aligned}$$

□

By using Lemma 5, Lemma 1 can be proved as follows.

Proof of Lemma 1. We have,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{I}\left\{U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}\right] \\
&= \sum_{t=K+1}^{T}\Pr\left\{U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\} \\
&\leq \sum_{t=K+1}^{T}\Pr\left\{U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}, \hat{\mu}_{i^*, N_{i^*}(t)} > \mu^*(\epsilon)\right\} \\
&\quad + \sum_{t=K+1}^{T}\Pr\left\{U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}, \hat{\tau}_{i^*, N_{i^*}(t)} < \tau^*(\epsilon)\right\} \\
&\quad + \sum_{t=K+1}^{T}\Pr\left\{U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}, \right. \\
&\qquad\qquad \left. \hat{\mu}_{i^*, N_{i^*}(t)} < \mu^*(\epsilon), \hat{\tau}_{i^*, N_{i^*}(t)} \geq \tau^*(\epsilon)\right\}.
\end{aligned}$$

The first summation is bounded by

$$\sum_{t=1}^{T}\Pr\left\{U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}, \hat{\mu}_{i^*, N_{i^*}(t)} > \mu^*(\epsilon)\right\}$$

$$
\begin{aligned}
&\leq \Pr \sum_{t=1}^{T} \Bigg\{ d_{\mathrm{KL}}\left(\hat{\tau}_{i^*,N_{i^*}(t)}, \tau^*(\epsilon)\right) > \frac{\ln(t)}{N_{i^*}(t)}, \\
&\qquad\qquad \hat{\tau}_{i^*,N_{i^*}(t)} \geq \tau^*(\epsilon) \Bigg\} \\
&\Bigg(\text{because } \frac{\hat{\mu}_{i^*,N_{i^*}(t)}}{\hat{\tau}_{i^*,N_{i^*}(t)}} \leq U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \\
&\qquad\qquad \Rightarrow \hat{\tau}_{i^*,N_{i^*}(t)} \geq \tau^*(\epsilon) \\
&\text{and } U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} < \frac{\hat{\mu}_{i^*,N_{i^*}(t)}}{\tau^*(\epsilon)} \\
&\quad \Rightarrow (\hat{\mu}_{i^*,N_{i^*}(t)}, \tau^*(\epsilon)) \\
&\qquad\qquad \notin \Phi\left(\hat{\mu}_{i^*,N_{i^*}(t)}, \hat{\tau}_{i^*,N_{i^*}(t)}, \frac{\ln(t)}{N_{i^*}(t)}\right) \Bigg) \\
&\leq \sum_{t=1}^{T} \sum_{n=1}^{t} \Pr\{ t < \exp\left(n d_{\mathrm{KL}}\left(\hat{\tau}_{i^*,n}, \tau^*(\epsilon)\right)\right), \\
&\qquad\qquad \hat{\tau}_{i^*,n} \geq \tau^*(\epsilon), N_{i^*}(t) = n \} \\
&= \sum_{n=1}^{T} \mathbb{E}\Bigg[\sum_{t=n}^{T} \mathbb{I}\{ t < \exp\left(n d_{\mathrm{KL}}\left(\hat{\tau}_{i^*,n}, \tau^*(\epsilon)\right)\right), \\
&\qquad\qquad N_{i^*}(t) = n \} \cdot \mathbb{I}\left\{\hat{\tau}_{i^*,n} \geq \tau^*(\epsilon)\right\} \Bigg] \\
&\leq \sum_{n=1}^{T} \mathbb{E}\left[\exp\left(n d_{\mathrm{KL}}\left(\hat{\tau}_{i^*,n}, \tau^*(\epsilon)\right)\right) \cdot \mathbb{I}\left\{\hat{\tau}_{i^*,n} \geq \tau^*(\epsilon)\right\}\right] \\
&= \sum_{n=1}^{T} \int_{y=1}^{\tau^*(\epsilon)} \exp\left(n d_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) \mathrm{d}F_n^{\tau}(y) \\
&= O(\epsilon^{-4}),
\end{aligned}
$$

where the last equality is by Lemma 5. In the same manner, we upper-bound the second summation as

$$
\sum_{t=1}^{T} \Pr\left\{ U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}, \hat{\tau}_{i^*,N_{i^*}(t)} < \tau^*(\epsilon) \right\} = O(\epsilon^{-4})
$$

and the third summation as

$$
\begin{aligned}
&\sum_{t=1}^{T} \Pr\Bigg\{ U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}, \\
&\qquad \hat{\mu}_{i^*,N_{i^*}(t)} \leq \mu^*(\epsilon), \hat{\tau}_{i^*,N_{i^*}(t)} \geq \tau^*(\epsilon) \Bigg\} \\
&\leq \sum_{t=1}^{T} \Pr\Bigg\{ d_{\mathrm{KL}}(\hat{\mu}_{i^*,N_{i^*}(t)}, \mu^*(\epsilon)) \\
&\qquad + d_{\mathrm{KL}}(\hat{\tau}_{i^*,N_{i^*}(t)}, \tau^*(\epsilon)) > \frac{\ln(t)}{N_{i^*}(t)}, \\
&\qquad \hat{\mu}_{i^*,N_{i^*}(t)} \leq \mu^*(\epsilon), \hat{\tau}_{i^*,N_{i^*}(t)} \geq \tau^*(\epsilon) \Bigg\} \\
&\Bigg(\text{because} \quad U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \\
&\Rightarrow \mu^*(\epsilon), \tau^*(\epsilon)) \notin \Phi\left(\hat{\mu}_{i^*,N_{i^*}(t)}, \hat{\tau}_{i^*,N_{i^*}(t)}, \frac{\ln(t)}{N_{i^*}(t)}\right) \Bigg) \\
&\leq \sum_{t=1}^{T} \sum_{n=1}^{t} \Pr\Bigg\{ t < \exp(n d_{\mathrm{KL}}(\hat{\mu}_{i^*,n}, \mu^*(\epsilon)) \\
&\qquad + n d_{\mathrm{KL}}(\hat{\tau}_{i^*,n}, \tau^*(\epsilon))), \\
&\qquad N_{i^*}(t) = n, \hat{\mu}_{i^*,n} \leq \mu^*(\epsilon), \hat{\tau}_{i^*,n} \geq \tau^*(\epsilon) \Bigg\} \\
&= \sum_{n=1}^{T} \mathbb{E}\Bigg[\sum_{t=n}^{T} \mathbb{I}\{ t < \exp\left(n d_{\mathrm{KL}}(\hat{\mu}_{i^*,n}, \mu^*(\epsilon))\right. \\
&\qquad \left. + n d_{\mathrm{KL}}(\hat{\tau}_{i^*,n}, \tau^*(\epsilon))\right), N_{i^*}(t) = n \} \\
&\qquad \cdot \mathbb{I}\{\hat{\mu}_{i^*,n} \leq \mu^*(\epsilon), \hat{\tau}_{i^*,n} \geq \tau^*(\epsilon)\} \Bigg] \\
&\leq \sum_{n=1}^{T} \mathbb{E}\Bigg[\exp(n d_{\mathrm{KL}}(\hat{\mu}_{i^*,n}, \mu^*(\epsilon)) \cdot \mathbb{I}\{\hat{\mu}_{i^*,n} \leq \mu^*(\epsilon)\} \\
&\qquad \cdot \exp(n d_{\mathrm{KL}}(\hat{\tau}_{i^*,n}, \tau^*(\epsilon)) \cdot \mathbb{I}\{\hat{\tau}_{i^*,n} \geq \tau^*(\epsilon)\} \Bigg] \\
&= \sum_{n=1}^{T} \int_{x=0}^{\mu^*(\epsilon)} \exp\left(n d_{\mathrm{KL}}\left(x, \mu^*(\epsilon)\right)\right) \mathrm{d}F_n^{\mu}(x) \\
&\qquad \cdot \int_{y=1}^{\tau^*(\epsilon)} \exp\left(n d_{\mathrm{KL}}\left(y, \tau^*(\epsilon)\right)\right) \mathrm{d}F_n^{\tau}(y) \\
&= O(\epsilon^{-6}),
\end{aligned}
$$

where the last inequality is by Lemma 5.

In summary, we obtain

$$
\sum_{t=K+1}^{T} \Pr\left\{ U_{i^*}(t) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \right\} = O(\epsilon^{-6}).
$$

□

Appendix C: Proof of Lemma 2

We bound $\sum_{t=K+1}^{T} \Pr\left\{ I(t) = i, U_i(t) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \right\}$ by some large number $\bar{N}_i(\epsilon)$ plus $1/\epsilon^2$. Lemma 6 states that, if $\bar{N}_i(\epsilon)$ is defined appropriately, it is bounded by $\frac{1+O(\epsilon)}{D_i} \ln(T)$.

Lemma 6. *Let $\epsilon > 0$ be sufficiently small and*

$$
\bar{N}_i(\epsilon) = \min\left\{ n : \phi\left(\mu_i + \epsilon, \tau_i - \epsilon, \frac{\ln(T)}{n}\right) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)} \right\}
$$

for any suboptimal action $i \neq i^*$. Then $\bar{N}_i(\epsilon)$ is bounded as:

$$
\bar{N}_i(\epsilon) = \frac{1 + O(\epsilon)}{D_i} \ln(T) + 1. \tag{A·5}
$$

Proof. Since $\phi\left(\mu_i + \epsilon, \tau_i - \epsilon, \frac{\ln(T)}{n}\right) < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}$ is equivalent to $(a\mu^*(\epsilon), a\tau^*(\epsilon)) \notin \Phi\left(\mu_i + \epsilon, \tau_i - \epsilon, \frac{\ln(T)}{n}\right)$ for all $a > 0$, $\bar{N}_i(\epsilon)$ is expressed as

$$
\begin{aligned}
\bar{N}_i(\epsilon) = \min\Bigg\{ n : \min_{a>0}\Big(& d_{\mathrm{KL}}\left(\mu_i + \epsilon, a\mu^*(\epsilon)\right) \\
& + d_{\mathrm{KL}}\left(\tau_i - \epsilon, a\tau^*(\epsilon)\right)\Big) > \frac{\ln(T)}{n} \Bigg\}.
\end{aligned}
$$

Because a that minimizes $d_{\mathrm{KL}}(\mu_i+\epsilon, a\mu^*(\epsilon)) + d_{\mathrm{KL}}(\tau_i-\epsilon, a\tau^*(\epsilon))$ lies[†] in an interval

$$A = \left[\frac{\mu_i+\epsilon}{\mu^*(\epsilon)}, \frac{\tau_i-\epsilon}{\tau^*(\epsilon)}\right],$$

there exists a constant h_i that satisfies[††]

$$d_{\mathrm{KL}}(\mu_i+\epsilon, a(\mu^*-\epsilon)) \geq d_{\mathrm{KL}}(\mu_i, a\mu^*) - h_i\epsilon$$

and

$$d_{\mathrm{KL}}(\tau_i-\epsilon, a(\tau^*+\epsilon)) \geq d_{\mathrm{KL}}(\tau_i, a\tau^*) - h_i\epsilon$$

for any appropriate a and sufficiently small ϵ. Thus,

$$\begin{aligned}
&\bar{N}_i(\epsilon)\\
&\leq \min\left\{n : \min_a \left(d_{\mathrm{KL}}(\mu_i, a\mu^*) + d_{\mathrm{KL}}(\tau_i, a\tau^*)\right)\right.\\
&\qquad\qquad \left. - 2h_i\epsilon > \frac{\ln(T)}{n}\right\}\\
&\leq \frac{\ln(T)}{D_i - 2h_i\epsilon} + 1 = \frac{\ln(T)}{D_i}\frac{1}{1-2h_i\epsilon D_i^{-1}} + 1\\
&\leq \frac{1+4h_i\epsilon D_i^{-1}}{D_i}\ln(T) + 1
\end{aligned}$$

holds by using the fact that $(1-x)^{-1} \leq (1+2x)$ for $x \in [0, 1/2]$. □

Now we prove Lemma 2 using Lemma 6.

Proof of Lemma 2. Since $I(t) = i \neq i^*$ implies $U_{i^*}(t) \leq U_i(t)$, we have

$$\begin{aligned}
&\sum_{t=K+1}^{T} \mathbb{I}\left\{I(t)=i, U_{i^*}(t) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}\\
&\leq \sum_{t=K+1}^{T} \mathbb{I}\left\{I(t)=i, U_i(t) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}\\
&= \sum_{t=K+1}^{T}\sum_{n=1}^{t} \mathbb{I}\left\{N_i(t)=n, I(t)=i,\right.\\
&\qquad\qquad \left.\phi\left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(t)}{n}\right) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}\\
&\leq \sum_{n=1}^{T}\sum_{t=n}^{T} \mathbb{I}\left\{N_i(t)=n, I(t)=i,\right.\\
&\qquad\qquad \left.\phi\left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}\right) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}\\
&\leq \sum_{n=1}^{T} \mathbb{I}\left\{\phi\left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}\right) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}\\
&\qquad (\{N_i(t)=n, I(t)=i\} \text{ occurs at most once}).
\end{aligned} \tag{A·6}$$

Moreover, (A·6) is transformed as:

$$\begin{aligned}
&\sum_{n=1}^{T} \mathbb{I}\left\{\phi\left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}\right) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\} \qquad (A·7)\\
&\leq \bar{N}_i(\epsilon) - 1\\
&\qquad + \sum_{n=\bar{N}_i(\epsilon)}^{T} \mathbb{I}\left\{\phi\left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}\right) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}\\
&\leq \bar{N}_i(\epsilon) - 1\\
&\quad + \sum_{n=\bar{N}_i(\epsilon)}^{\infty} (\mathbb{I}\{\hat{\mu}_{i,n} \geq \mu_i+\epsilon\} + \mathbb{I}\{\hat{\tau}_{i,n} \leq \tau_i-\epsilon\}),
\end{aligned} \tag{A·8}$$

where the last transformation is derived by using the fact that $\{\hat{\mu}_{i,n} < \mu_i+\epsilon, \hat{\tau}_{i,n} > \tau_i-\epsilon, n \geq \bar{N}_i(\epsilon)\}$ implies

$$\begin{aligned}
\phi\left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}\right) &\leq \phi\left(\mu_i+\epsilon, \tau_i-\epsilon, \frac{\ln(T)}{n}\right)\\
&< \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}.
\end{aligned}$$

Taking the expectation of (A·8) yields

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{I(t)=i, U_{i^*}(t) \geq \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}\right\}\right]\\
&\leq \bar{N}_i(\epsilon) - 1 + \mathbb{E}\left[\sum_{n=\bar{N}_i(\epsilon)}^{\infty} (\mathbb{I}\{\hat{\mu}_{i,n} \geq \mu_i+\epsilon\}\right.\\
&\qquad\qquad \left. + \mathbb{I}\{\hat{\tau}_{i,n} \leq \tau_i-\epsilon\})\right]\\
&\leq \bar{N}_i(\epsilon) - 1 + 2\sum_{n=1}^{\infty} \exp(-2n\epsilon^2)\\
&\leq \bar{N}_i(\epsilon) - 1 + 2\int_0^{\infty} \exp(-2\epsilon^2 x)\mathrm{d}x\\
&\leq \bar{N}_i(\epsilon) - 1 + \frac{1}{\epsilon^2},
\end{aligned}$$

where we used Hoeffding's inequality (Fact 3). Using Lemma 6 completes the proof. □

Appendix D: Proof of Theorem 3

The following lemma is used in the proof of Theorem 3.

[†]Let $f_{\alpha,\beta}(a) = d_{\mathrm{KL}}(\alpha, \beta a)$. Then, $f_{\alpha,\beta}(a)$ is decreasing for $a < \frac{\alpha}{\beta}$ and increasing for $a > \frac{\alpha}{\beta}$. Using this fact, we can know that $f_{\mu_i+\epsilon,\mu^*(\epsilon)}(a) + f_{\tau_i-\epsilon,\tau^*(\epsilon)}(a)$ is decreasing for $a < \frac{\mu_i+\epsilon}{\mu^*(\epsilon)}$ and increasing for $a > \frac{\tau_i-\epsilon}{\tau^*(\epsilon)}$, which means a that minimizes $f_{\mu_i+\epsilon,\mu^*(\epsilon)}(a) + f_{\tau_i-\epsilon,\tau^*(\epsilon)}(a)$ lies in the interval A. Note that $\frac{\mu_i+\epsilon}{\mu^*(\epsilon)} < \frac{\tau_i-\epsilon}{\tau^*(\epsilon)}$ holds for sufficiently small $\epsilon > 0$ because $\frac{\mu_i+\epsilon}{\tau_i-\epsilon} < \frac{\mu^*(\epsilon)}{\tau^*(\epsilon)}$ holds for sufficiently small $\epsilon > 0$ by the oprtimality of (μ^*, τ^*).

[††]Let $g_1(\epsilon, a) = d_{\mathrm{KL}}(\mu_i+\epsilon, a\mu^*(\epsilon))$, $g_2(\epsilon, a) = d_{\mathrm{KL}}(\tau_i-\epsilon, a\tau^*(\epsilon))$. Then,

$$h_i = \sup_{a\in A, \epsilon\leq\epsilon_0} \max\left\{\frac{\partial g_1(\epsilon,a)}{\partial\epsilon}, \frac{\partial g_2(\epsilon,a)}{\partial\epsilon}\right\},$$

which is bounded for sufficiently small $\epsilon_0 > 0$, is enough for all $\epsilon \leq \epsilon_0$.

Lemma 7 (Maximal law of large numbers [23]). *Let $X_1, \ldots$ be i.i.d. random variables with positive mean. if $\frac{1}{n}\sum_{t=1}^{n} X_t$ converges to μ almost surely as $n \to \infty$, then*

$$\frac{1}{n}\max_{s\le n}\sum_{t=1}^{s} X_t$$

also converges to μ almost surely as $n \to \infty$.

Proof of Theorem 3. For any suboptimal action i and small positive constant $\epsilon > 0$, there exist distribution parameters μ_i' and τ_i' such that $\mu_i' \ge \mu_i$, $\tau_i' < \tau_i$,

$$\frac{\mu_i'}{\tau_i'} > \frac{\mu^*}{\tau^*} > \frac{\mu_i}{\tau_i},$$

and

$$\begin{aligned} &d_{\mathrm{KL}}(\mu_i, \mu_i') + d_{\mathrm{KL}}(\tau_i, \tau_i') \\ &\le (1+\epsilon)\min_{a>0}\left(d_{\mathrm{KL}}(\mu_i, a\mu^*) + d_{\mathrm{KL}}(\tau_i, a\tau^*)\right) \\ &= (1+\epsilon)D_i. \end{aligned} \tag{A·9}$$

In the following, we consider a modified bandit problem such that the parameters of action i is not (τ_i, μ_i) but (τ_i', μ_i'). Note that, unlike the original bandit problem, in the modified bandit problem the optimal action is not action i^* but action i.

Let $p_X(x)$ and $p_C(x)$ be the probability density functions (PDFs) of Bernoulli distributions $\mathcal{B}(\mu_i)$ and $\mathcal{B}(\tau_i)$ respectively. Moreover, let $q_X(x)$ and $q_C(x)$ be the PDFs of $\mathcal{B}(\mu_i')$ and $\mathcal{B}(\tau_i')$ respectively.

For the sequence of random variables $\{(X_{i,s}, C_{i,s})\}_{s=1}^{T}$, where $\{X_{i,s}\}$ and $\{C_{i,s}\}$ are the i.i.d. samples from $\mathcal{B}(\mu_i)$ from $\mathcal{B}(\tau_i)$ respectively, define the *empirical log-likelihood ratio* function

$$\begin{aligned} &\mathrm{ELLR}(n) \\ &= \ln\left(\frac{p_X(X_{i,1})\cdots p_X(X_{i,n})\cdot p_C(C_{i,1})\cdots p_C(C_{i,n})}{q_X(X_{i,1})\cdots q_X(X_{i,n})\cdot q_C(C_{i,1})\cdots q_C(C_{i,n})}\right) \\ &= \sum_{s=1}^{n}\left(\ln\frac{p_X(X_{i,s})}{q_X(X_{i,s})} + \ln\frac{p_C(C_{i,s})}{q_C(C_{i,s})}\right) \end{aligned}$$

Note that

$$\begin{aligned} &\mathbb{E}\left[\ln\frac{p_X(X_{i,s})}{q_X(X_{i,s})} + \ln\frac{p_C(C_{i,s})}{q_C(C_{i,s})}\right] \\ &\qquad = d_{\mathrm{KL}}(\mu_i, \mu_i') + d_{\mathrm{KL}}(\tau_i, \tau_i') \end{aligned}$$

holds.

Let $\Pr'$ and $\mathbb{E}'$ be the probability and expectation with respect to the modified bandit problem. For any event E_T,

$$\Pr'\{E_T\} = \mathbb{E}\left[\mathbb{I}\{E_T\}\exp\left(-\mathrm{ELLR}(N_i(T))\right)\right] \tag{A·10}$$

holds.

Define two events

$$A_T = \{N_i(T) \le S(T)\}$$

and

$$B_T = \{\mathrm{ELLR}(N_i(T)) \le (1-\epsilon')\ln(t)\}$$

where $\epsilon' \in (0, \epsilon)$ and

$$S(T) = \frac{(1-\epsilon)\ln(T)}{d_{\mathrm{KL}}(\mu_i, \mu_i') + d_{\mathrm{KL}}(\tau_i, \tau_i')}.$$

In the following, we first derive $\Pr\{A_T\} = o(1)$ and then a lower-bound of $\mathbb{E}[N_i(T)]$.

By using Eq. (A·10) and the definition of A_T and B_T, we have

$$\begin{aligned} \Pr'\{A_T, B_T\} &= \mathbb{E}[\mathbb{I}\{A_T, B_T\}\exp(-\mathrm{ELLR}(N_i(T)))] \\ &\ge \mathbb{E}[\mathbb{I}\{A_T, B_T\}\exp(-(1-\epsilon')\ln(T))] \\ &= \Pr\{A_T, B_T\}T^{-(1-\epsilon')}. \end{aligned}$$

Then,

$$\begin{aligned} \Pr\{A_T, B_T\} &\le T^{1-\epsilon'}\Pr'\{A_T, B_T\} \\ &\le T^{1-\epsilon'}\Pr'\{N_i(T) \le S(T)\} \\ &= T^{1-\epsilon'}\Pr'\{T - N_i(T) \ge T - S(T)\}. \end{aligned}$$

By using the Markov's inequality (Fact 1), we obtain

$$\Pr\{A_T, B_T\} \le T^{1-\epsilon'}\frac{\mathbb{E}'[T - N_i(T)]}{T - S(T)}.$$

On the modified bandit problem, the optimal action is not i^* but i. Therefore, the strong consistency of the policy requires $\mathbb{E}'[T - N_i(T)] = \mathbb{E}'[\sum_{j\ne i} N_j(T)] = o(T^a)$ for any $a > 0$. By using this we obtain

$$\Pr\{A_T, B_T\} = o(1). \tag{A·11}$$

Moreover, let $\bar{B}_T$ be the complement event of B_T. We have

$$\begin{aligned} &\Pr\{A_T, \bar{B}_T\} \\ &= \Pr\left\{N_i(T) \le S(T), \mathrm{ELLR}(N_i(T)) > (1-\epsilon')\ln(T)\right\} \\ &\le \Pr\left\{\max_{s\le S(T)}\mathrm{ELLR}(s) > (1-\epsilon')\ln(T)\right\} \\ &= \Pr\left\{\frac{1}{S(T)}\max_{s\le S(T)}\mathrm{ELLR}(s)\right. \\ &\qquad \left. > \frac{1-\epsilon'}{1-\epsilon}(d_{\mathrm{KL}}(\mu_i, \mu_i') + d_{\mathrm{KL}}(\tau_i, \tau_i'))\right\} \end{aligned}$$

Using Lemma 7, and the expectation of $\mathrm{ELLR}(s)$,

$$\begin{aligned} &\lim_{T\to\infty}\frac{1}{S(T)}\max_{s\le S(T)}\mathrm{ELLR}(s) \\ &\qquad \to d_{\mathrm{KL}}(\mu_i, \mu_i') + d_{\mathrm{KL}}(\tau_i, \tau_i') \quad \text{a.s.} \end{aligned}$$

and, using the fact that $(1-\epsilon')/(1-\epsilon) > 1$, we obtain

$$\Pr\{A_T, \bar{B}_T\} = o(1). \tag{A·12}$$

By (A·11) and (A·12), we have

$$\Pr\{A_T\} = \Pr\{A_T, B_T\} + \Pr\{A_T, \bar{B}_T\} = o(1). \quad (\text{A}\cdot 13)$$

Finally, $\mathbb{E}[N_i(T)]$ is lower-bounded as:

$$\begin{aligned}
&\mathbb{E}[N_i(T)] \\
&\geq S(T) \cdot \Pr\{N_i(T) \geq S(T)\} \\
&\qquad \text{(by Markov's inequality (Fact 1))} \\
&= S(T)(1 - \Pr\{A_T\}) \\
&= \frac{(1-\epsilon)\ln(T)}{d_{\mathrm{KL}}(\mu_i, \mu_i') + d_{\mathrm{KL}}(\tau_i, \tau_i')}(1 - o(1)) \quad \text{(by (A}\cdot 13\text{))} \\
&\geq \frac{1-\epsilon}{1+\epsilon}(1 - o(1))\frac{\ln(T)}{D_i} \quad \text{(by (A}\cdot 9\text{))} \\
&\geq (1 - o(1))\frac{\ln(T)}{D_i},
\end{aligned}$$

where the last inequality holds by choosing sufficiently small $\epsilon > 0$ that converges to 0 when $T \to \infty$. □

Appendix E: Proof of Lemma 3

Proof of Lemma 3. Let $C_{i,s}$ denote the cost $C_i(t)$ for the player's s-th choice of action i. We first derive Ineq. (6). For $T = 1, 2, \cdots$, let $z = \frac{\tau_{\min} T}{BK}$. Then, we have

$$\begin{aligned}
&\Pr\{T(B) \geq T\} = \Pr\left\{T(B) \geq \frac{zBK}{\tau_{\min}}\right\} \\
&\leq \Pr\left\{\bigcup_{i\in[K]}\left\{\sum_{s=1}^{\left\lceil \frac{zB}{\tau_{\min}}\right\rceil} C_{i,s} \leq B\right\}\right\} \\
&\qquad \left(\text{because } N_i\left(\frac{zBK}{\tau_{\min}}\right) \geq \left\lceil \frac{zB}{\tau_{\min}}\right\rceil \text{ and } \sum_{s=1}^{N_i(zBK/\tau_{\min})} C_{i,s} \leq B \text{ for some } i\right) \\
&\leq \sum_{i=1}^{K} \Pr\left\{\sum_{s=1}^{\left\lceil \frac{zB}{\tau_{\min}}\right\rceil} C_{i,s} \leq B\right\} \quad \text{(by the union bound)} \\
&\leq \sum_{i=1}^{K} \Pr\left\{\frac{1}{\left\lceil \frac{zB}{\tau_{\min}}\right\rceil}\sum_{s=1}^{\left\lceil \frac{zB}{\tau_{\min}}\right\rceil} C_{i,s} \leq \tau_i - \frac{z-1}{z}\tau_{\min}\right\} \\
&\qquad \left(\text{by } \frac{1}{\left\lceil \frac{zB}{\tau_{\min}}\right\rceil} \leq \frac{1}{\frac{zB}{\tau_{\min}}} \text{ and } \tau_i \geq \tau_{\min}\right) \\
&\leq K\mathrm{e}^{-2\left(\frac{zB}{\tau_{\min}}\right)((z-1)/z)^2\tau_{\min}^2} \\
&\qquad \text{(by the Hoeffding inequality)} \\
&= K\mathrm{e}^{-2((z-1)^2/z)\tau_{\min}B} \quad (\text{A}\cdot 14) \\
&\leq K\mathrm{e}^{-2\tau_{\min}B(z-2)}. \quad (\text{A}\cdot 15)
\end{aligned}$$

For any action i,

$$\begin{aligned}
&\mathbb{E}[N_i(T(B))] \\
&= \sum_{T=1}^{\infty} \mathbb{E}[N_i(T)] \Pr\{T(B) = T\} \\
&\leq \sum_{T=0}^{\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor} \mathbb{E}[N_i(T)] \Pr\{T(B) = T\} \\
&\quad + \sum_{\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor+1}^{\infty} T \Pr\{T(B) = T\} \\
&\text{(because } \mathbb{E}[N_i(T)] \leq T\text{)} \\
&\leq \sum_{T=0}^{\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor} \mathbb{E}\left[N_i\left(\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] \Pr\{T(B) = T\} \\
&\quad + \sum_{T=\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor+1}^{\infty} T(\Pr\{T(B) \geq T\} \\
&\qquad\qquad - \Pr\{T(B) \geq T+1\}) \\
&\leq \mathbb{E}\left[N_i\left(\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] \\
&\quad + \left(\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor + 1\right) \Pr\left\{T(B) \geq \left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor + 1\right\} \\
&\quad + \sum_{T=\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor+2}^{\infty} \Pr\{T(B) \geq T\} \\
&\leq \mathbb{E}\left[N_i\left(\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] + \left(\frac{2BK}{\tau_{\min}} + 1\right) K\mathrm{e}^{-\tau_{\min}B} \\
&\quad + \sum_{T=\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor+2}^{\infty} \Pr\{T(B) \geq T\} \\
&\left(\text{by } (\text{A}\cdot 14)\right. \\
&\qquad \left.\text{and } z = \frac{\tau_{\min}T}{BK} \geq 2 \text{ for } T = \left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor + 1\right) \\
&\leq \mathbb{E}\left[N_i\left(\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] + o(1) \\
&\quad + \sum_{T=\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor+2}^{\infty} K\mathrm{e}^{-2\tau_{\min}B\left(\frac{\tau_{\min}T}{BK}-2\right)} \quad \text{(by (A}\cdot 15\text{))} \\
&\leq \mathbb{E}\left[N_i\left(\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] + o(1) \\
&\qquad + \frac{K\mathrm{e}^{-2\frac{\tau_{\min}^2}{K}\left(\left\lfloor \frac{2BK}{\tau_{\min}}\right\rfloor+2\right)}\mathrm{e}^{4\tau_{\min}B}}{1 - \mathrm{e}^{-2\frac{\tau_{\min}^2}{K}}} \\
&\qquad \text{(by the sum of geometrical series)}
\end{aligned}$$

$$\leq \mathbb{E}\left[N_i\left(\left\lfloor\frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] + o(1) + \frac{K\mathrm{e}^{-2\frac{\tau_{\min}^2}{K}}}{1-\mathrm{e}^{-2\frac{\tau_{\min}^2}{K}}}$$
$$\left(\text{using } \mathrm{e}^{-2\frac{\tau_{\min}^2}{K}\left(\left\lfloor\frac{2BK}{\tau_{\min}}\right\rfloor+2\right)} \leq \mathrm{e}^{-2\frac{\tau_{\min}^2}{K}\left(\frac{2BK}{\tau_{\min}}+1\right)}\right)$$
$$\leq \mathbb{E}\left[N_i\left(\left\lfloor\frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] + o(1) + \frac{K^2}{2\tau_{\min}^2}$$
$$\left(\text{by } \frac{\mathrm{e}^{-x}}{1-\mathrm{e}^{-x}} = \frac{1}{\mathrm{e}^x-1} \leq \frac{1}{x} \text{ for } x \geq 0\right)$$

Then we find that

$$\begin{aligned}
&\tilde{R}^\pi(B)\\
&= \sum_{i:\,\Delta_i>0} \tau_i\Delta_i\,\mathbb{E}[N_i(T(B))]\\
&\leq \sum_{i:\,\Delta_i>0} \tau_i\Delta_i\left(\mathbb{E}\left[N_i\left(\left\lfloor\frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] + \frac{K^2}{2\tau_{\min}^2} + o(1)\right)\\
&= \sum_{i:\,\Delta_i>0} \tau_i\Delta_i\left(\mathbb{E}\left[N_i\left(\left\lfloor\frac{2BK}{\tau_{\min}}\right\rfloor\right)\right] + \frac{K^2}{2\tau_{\min}^2}\right) + o(1).
\end{aligned}$$

for any policy.

Next we derive Ineq. (7). First, we have

$$\begin{aligned}
&\Pr\left\{T(B) \leq \left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor\right\}\\
&\leq \Pr\left\{\bigcup_{i\in[K]}\left\{\sum_{s=1}^{\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor} C_{i,s} \geq \frac{B}{K}\right\}\right\}\\
&\quad\left(\text{because } \sum_{s=1}^{N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor\right)} C_{i,s} \geq \frac{B}{K} \text{ for some } i\right)\\
&\leq \sum_{i=1}^{K}\Pr\left\{\sum_{s=1}^{\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor} C_{i,s} \geq \frac{B}{K}\right\} \quad \text{(by the union bound)}\\
&\leq \sum_{i=1}^{K}\Pr\left\{\frac{1}{\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor}\sum_{s=1}^{\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor} C_{i,s} \geq \tau_i + \tau_{\max}\right\}\\
&\quad\left(\text{by } \frac{1}{\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor} \geq \frac{1}{\frac{B}{2K\tau_{\max}}} \text{ and } \tau_i \leq \tau_{\max}\right)\\
&\leq K\mathrm{e}^{-2\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor\tau_{\max}^2} \quad \text{(by the Hoeffding inequality)}\\
&\leq K\mathrm{e}^{-2\left(\frac{B}{2K\tau_{\max}}-1\right)\tau_{\max}^2} \qquad \text{(A·16)}
\end{aligned}$$

For any action i,

$$\begin{aligned}
&\mathbb{E}[N_i(T(B))]\\
&= \sum_{T=1}^{\infty}\mathbb{E}[N_i(T)]\Pr\{T(B)=T\}\\
&\geq \sum_{T=\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor+1}^{\infty}\mathbb{E}[N_i(T)]\Pr\{T(B)=T\}\\
&\geq \mathbb{E}\left[N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor+1\right)\right]\\
&\qquad\cdot\left(1-\Pr\left\{T(B)\leq\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor\right\}\right)\\
&\geq \mathbb{E}\left[N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor+1\right)\right]\\
&\qquad\cdot\left(1-K\mathrm{e}^{-2\left(\frac{B}{2K\tau_{\max}}-1\right)\tau_{\max}^2}\right) \text{ (by A·16)}\\
&\geq \mathbb{E}\left[N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor+1\right)\right]\\
&\qquad-\left(\frac{B}{2\tau_{\max}}+K\right)\mathrm{e}^{-2\left(\frac{B}{2K\tau_{\max}}-1\right)\tau_{\max}^2}\\
&\quad\left(\text{by } \mathbb{E}\left[N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor+1\right)\right] \leq \frac{B}{2K\tau_{\max}}+1\right)\\
&\geq \mathbb{E}\left[N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor+1\right)\right] - o(1)
\end{aligned}$$

Then we find that

$$\begin{aligned}
&\tilde{R}^\pi(B)\\
&= \sum_{i:\,\Delta_i>0}\tau_i\Delta_i\,\mathbb{E}[N_i(T(B))]\\
&\geq \sum_{i:\,\Delta_i>0}\tau_i\Delta_i\left(\mathbb{E}\left[N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor+1\right)\right] - o(1)\right)\\
&= \sum_{i:\,\Delta_i>0}\tau_i\Delta_i\,\mathbb{E}\left[N_i\left(\left\lfloor\frac{B}{2K\tau_{\max}}\right\rfloor+1\right)\right] - o(1).
\end{aligned}$$

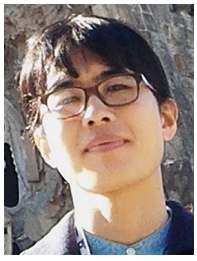
□

Appendix F: Distribution Parameters Used in Experiments on Real-World Dataset

Table A·1 Reward and cost distributions based on real-world dataset.

i	$X_i(t)$	$C_i(t)$	τ_i	$\frac{\mu_i}{\tau_i}$
1	$\mathcal{B}(0.0379)$	$\Lambda(4.13, 0.837^2)$	88.6	0.000428
2	$\mathcal{B}(0.0220)$	$\Lambda(3.80, 0.900^2)$	67.3	0.000327
3	$\mathcal{B}(0.0218)$	$\Lambda(4.02, 0.903^2)$	83.5	0.000262
4	$\mathcal{B}(0.0376)$	$\Lambda(4.11, 0.698^2)$	77.6	0.000485
5	$\mathcal{B}(0.0406)$	$\Lambda(4.15, 0.745^2)$	84.0	0.000483
6	$\mathcal{B}(0.0403)$	$\Lambda(3.92, 0.928^2)$	77.3	0.000522
7	$\mathcal{B}(0.0501)$	$\Lambda(3.64, 0.951^2)$	59.7	0.000840
8	$\mathcal{B}(0.0554)$	$\Lambda(3.88, 0.868^2)$	70.5	0.000787
9	$\mathcal{B}(0.0206)$	$\Lambda(4.07, 0.697^2)$	74.9	0.000274
10	$\mathcal{B}(0.0125)$	$\Lambda(4.07, 0.628^2)$	71.3	0.000176
11	$\mathcal{B}(0.0259)$	$\Lambda(3.93, 0.910^2)$	77.3	0.000335
12	$\mathcal{B}(0.0391)$	$\Lambda(4.35, 0.742^2)$	102	0.000382
13	$\mathcal{B}(0.0459)$	$\Lambda(4.07, 1.06^2)$	102	0.000449
14	$\mathcal{B}(0.0220)$	$\Lambda(4.07, 0.633^2)$	71.6	0.000307
15	$\mathcal{B}(0.0250)$	$\Lambda(3.73, 0.997^2)$	68.3	0.000366
16	$\mathcal{B}(0.0305)$	$\Lambda(3.69, 0.985^2)$	64.7	0.000472
17	$\mathcal{B}(0.0623)$	$\Lambda(3.70, 1.07^2)$	72.0	0.000866
18	$\mathcal{B}(0.0253)$	$\Lambda(4.07, 0.914^2)$	88.7	0.000286
19	$\mathcal{B}(0.0412)$	$\Lambda(4.11, 0.525^2)$	70.2	0.000587
20	$\mathcal{B}(0.0103)$	$\Lambda(4.11, 0.577^2)$	72.3	0.000143

Ryo Watanabe received his M.S. in information science from Hokkaido University in 2014. He is currently a doctor course student in Graduate School of Information Science and Technology, Hokkaido University.



Junpei Komiyama received his M.S. in applied physics from the University of Tokyo in 2009. From 2009 to 2012, he worked for Dwango, Co., Ltd., where he was engaged in the development of a large scale streaming service. He received a D.S. degree in information science from the University of Tokyo in 2016. He has been a research associate at the University of Tokyo since 2016. His research interests have been in the area of machine learning and data mining, especially in online learning and pattern mining algorithms.

Atsuyoshi Nakamura received his M.S. and D.S. degrees in computer science from the Tokyo Institute of Technology in 1988 and 2000. From 1988 to 2002, he worked for NEC Corporation, where he was engaged in development of database management systems, research in machine learning and its application to the WWW. He has been an associate professor at Hokkaido University since 2002. His research interests have been in the area of machine learning and data mining, especially computational learning theory, information filtering, web mining and string mining.

Mineichi Kudo received his Dr. Eng. degree in Information Engineering from the Hokkaido University in 1988. Starting from an instructor, since 2001, he is a a professor (2001–) in Hokkaido University. In 2001 he received with professor Jack Sklansky "the twenty-seventh annual pattern recognition society award. He was elected to a fellow of the International Association for Pattern Recognition on December 10, 2008. His current research interests include design of pattern recognition systems, image processing, data mining and computational learning theory. He is a member of the Pattern Recognition Society and the IEEE.