



HOKKAIDO UNIVERSITY

Title	Analysis of Policies for Budgeted Multi-Armed Bandit Problem and Matching-Selection Bandit Problem
Author(s)	渡邊, 僚
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第13080号
Issue Date	2018-03-22
DOI	https://doi.org/10.14943/doctoral.k13080
Doc URL	https://hdl.handle.net/2115/70411
Type	doctoral thesis
File Information	Ryo_Watanabe.pdf



Doctorial Thesis

Analysis of the Policies
for Budgeted Multi-Armed Bandit Problem
and Matching-Selection Bandit Problem

Ryo Watanabe

Graduate School of Information Science and Technology
Hokkaido University, Japan

March, 2018

本論文の要旨

多腕バンディット問題とは、与えられた行動の選択肢から一つを選び、対応した報酬を得るというラウンドを繰り返す設定において、報酬の合計（累積報酬）を最大化する問題である。ただし、選んだ行動に対する報酬しか情報はフィードバックされず、他の行動を選んでいたら報酬はいくらであったかは明らかにされない。そのため、情報収集のための探索とそれまでに得られている情報を使った知識利用をバランス良く行う必要がある問題として知られている。この問題において、ある回数ラウンドを繰り返す場合に、累積報酬を最大化する選択ルール（方策）を設計する研究がなされてきた。

方策の設計は、報酬がどのように生成されるかに大きく影響される。大きく分けて確率的に生成されるという設定と、方策を知っている敵が選ぶという設定がある。本研究ではより応用範囲の広い確率的な設定のみ扱う。また、累積報酬そのものは問題依存であり、方策の評価指標として適切でないため、最適な方策（期待累積報酬最大の方策）との期待累積報酬の差であるリグレットで方策を評価する。

多腕バンディット問題の方策として、今までに様々な方策が提案され、そのリグレット解析がなされてきた。主なものとしては、UCB (Upper Confidence Bound) 方策や KL-UCB (Kullback-Leibler UCB) 方策 のような期待報酬の信頼上界に基づくもの、最適ではない行動の誤選択率を直接制御する DMED (Deterministic Minimum Empirical Divergence) 方策、事後確率分布に基づいてサンプリングした値で行動を評価する Thompson Sampling が知られており、それぞれリグレットの上界が証明されている。

一方、バンディット問題における理論限界、つまり厳密なリグレット下界も証明されている。実際、上で主なものとして挙げた4つの方策のうち UCB 方策を除く3つの方策は、それらのリグレット上界の支配項はラウンド数 T の関数としてみた場合、下界の支配項と漸近的に係数まで一致する。リグレット上界が下界と漸近的に一致する方策は、漸近最適方策 (Asymptotically Optimal Policy) と呼ばれており、KL-UCB 方策、DMED 方策及び Thompson Sampling は漸近最適方策であると言える。

本論文では、多腕バンディット問題の一般化である「予算制約付き多腕バンディット問題」及び「マッチング選択バンディット問題」に関して、主にリグレット解析を行う。予算制約付き多腕バンディット問題では、2つの

方策を提案し、それらの有効性をリグレット解析及びシミュレーション実験により示す。また、マッチング選択バンディット問題に関しては、既存の方策である LLR (Learning with Linear Rewards) 方策のリグレット上界を改善する。

予算制約つき多腕バンディット問題とは、行動の選択に際しコストが発生し、コスト総和が決められた予算を超えるまでのあいだ、ラウンドを繰り返すことのできるバンディット問題である。本論文では、行動毎にコストが異なるのみでなく、それが未知の確率分布に従って確率的に定まる場合を扱う。このバンディット問題に対するリグレット下界を示すと共に、KL-UCB 方策と UCB 方策をこの問題設定に合うようにそれぞれ拡張した KL-UCB-SC (KL-UCB for Stochastic Costs) 方策及び UCB-SC (UCB for Stochastic Costs) 方策を提案し、リグレット上界を求め、シミュレーション実験により有効性を示す。

KL-UCB 方策が予算制約のない普通のバンディット問題の漸近最適方策であるように、KL-UCB-SC 方策もまた、予算制約付き多腕バンディット問題の漸近最適方策であることを、リグレット上界を分析することにより示す。KL-UCB-SC 方策はリグレット評価においては高性能であるが、各々の行動に対する選択指標値を計算するのに最適化問題を解かなければならず、計算効率のよい実装が難しいという問題点がある。一方、UCB-SC 方策は、漸近最適方策であることを証明できないものの、各々の行動に対する選択指標値の計算式が閉形式で書けるため、計算効率の良い実装が可能である。シミュレーション実験では、これらの理論評価を裏付ける結果が得られたほか、KL-UCB-SC の変種である KL-UCB-SC+ は、既存手法の中で最も性能が良いとされている BTS (Budgeted Thompson Sampling) と同等の性能を示した。また、UCB-SC はこれら 2 つの方策よりかは実験によるリグレット性能も劣るものの、計算効率はこれらより良く、変種の UCB-SC+ は計算効率の良い他の方策 (PD-BwK, UCB-BV1) よりも良いリグレット性能を示した。

マッチング選択バンディット問題は、各ラウンドにおいて完全 2 部グラフにおける最大マッチング (共通の端点を持たない最大辺集合) の 1 つを選択し、選択した辺集合に対応する報酬を得る設定の問題である。この問題は、全ての辺の集合から、条件を満たす辺集合を選択する組み合わせバンディット問題の一種である。報酬は辺毎に発生し、選択した辺集合の辺各々から報酬が得られるという半バンディット (semi-bandit) 設定を扱う。本論文では、 $M \leq N$ を満たす自然数 M, N に対し、完全 2 部グラフ $K_{M,N}$ 上のマッチング選択バンディット問題において、既存方策である LLR 方策のリグレット上界を $\Theta(M^{2/3})$ 倍改善する。

Contents

1	Introduction	1
2	The Classical MAB Problem	5
2.1	Problem Formulation	5
2.1.1	Definition	5
2.1.2	Cumulative Expected Pseudo-Regret	6
2.2	Policies	7
2.2.1	The UCB (Upper Confidence Bound)	7
2.2.2	KL-UCB (Kullback-Leibler UCB)	8
2.3	Regret Analysis of UCB Family	9
2.3.1	Known Facts	9
2.3.2	Regret Analysis of the UCB	10
2.3.3	Regret Analysis of KL-UCB	13
2.4	Lower Bound of Regret	14
3	Budgeted Multi-Armed Bandit Problem	16
3.1	Problem Formulation	16
3.1.1	Definition	16
3.1.2	Considering a Stopping Time on Regret	17
3.2	Related Works	24
3.3	Policies	24
3.3.1	KL-UCB-SC	24
3.3.2	UCB-SC	25
3.4	Analytic Results	27
3.4.1	Dissimilarity of an Action between the Best Action	27
3.4.2	Upper Bound of the Expected Regret for KL-UCB-SC	28
3.4.3	Upper Bound of the Expected Regret for UCB-SC	37
3.5	Lower Bound of Regret	38
3.5.1	Comparison with other policies	42
3.6	Experiments	43
3.6.1	Setup	43

<i>CONTENTS</i>	iv
3.6.2 Results	44
3.7 Summary	45
4 Matching-Selection Bandit Problem	47
4.1 Problem Definition	47
4.2 LLR (Learning with Linear Rewards)	49
4.3 Regret Analysis	50
4.4 Summary	57
5 Conclusion	58

List of Tables

3.1	Pairs $(X_i(t), Y_i(t))$ of distributions for actions i	44
3.2	Computation time $[\mu\text{s}]$ per round averaged over 100 runs. . . .	45

List of Figures

3.1	The empirical regret curves averaged over 100 independent runs with $B = 10\,000$ for scenarios M and H.	46
-----	--	----

List of Algorithms

1	The Classical Multi-Armed Bandits	6
2	Budgeted Multi-Armed Bandits	17
3	Matching-Selection Bandits	48

Chapter 1

Introduction

Suppose the situation that repeats selection of an action from a given set of actions, and receives a reward corresponding to the selected action without obtaining any information about the other actions. In this situation, the knowledge available for choosing a current action is limited to the past selections and rewards. Here, we call such a situation as “multi-armed bandits” or “bandit environment”. In the multi-armed bandits, a rule for choosing next action, based on the knowledge, is called *policy*.

The classical multi-armed bandit problem is a problem (or one-player game) on the multi-armed bandits, in which rewards are generated according to action-specific probabilistic distributions, and a player attempts to maximize his/her cumulative reward over the whole trials [1–3]. The classical multi-armed bandit problem is just like a problem of the gambler who faced a number of slot machines in a casino. A slot machine is sometimes called as “one-armed bandit”, since the conventional mechanical slot machine has a lever (arm) to spin the reel, and it mostly robs the wealth of the player. This is why such a problem is called the “multi-armed bandit problem”. Historically, the classical multi-armed bandit problem was firstly formulated. After that, a problem without the stochastic assumption on the reward (e.g. adversarial bandits [4,5]) and a problem with different goal (e.g. best arm identification [6]) have been formulated. These problems are named with a word “bandit” since they can be seen as extensions, generalizations or variants of the classical multi-armed bandit problem. The situation called “multi-armed bandits” here is the common essence of all the problem which are named “bandit”.

Nowadays, several policies are designed for the classical multi-armed bandit problem. UCB (Upper Confidence Bound) family, such as the UCB (Auer’s UCB1 [3]) or KL-UCB [7,8] (Kullback-Leibler UCB) are based on an optimistic estimation of an expected reward for each action. DMED [9,10]

(Deterministic Empirical Minimum Divergence) maintains a candidate of best action. Thompson Sampling [1, 11] is usually implemented as a randomized algorithm.

As a quantitative evaluation of a policy for the multi-armed bandit problem, expected cumulative pseudo-regret is used. Cumulative pseudo-regret is defined as a difference between cumulative reward gained by the best policy and that gained by a given policy. The best policy is the policy who maximizes his/her expected cumulative reward, according to the hidden reward distribution. On the classical multi-armed bandit problem, a policy who chooses the best action regardless to the past is the best policy, where the best action is an action whose expected reward is highest. The upper bounds of expected cumulative pseudo-regret of above-mentioned four policies (UCB, KL-UCB, DMED, Thompson Sampling) are theoretically proved [3, 7, 11].

On the other hand, one may be curious how small we can make the expected cumulative pseudo-regret. In general, expected regret cannot be zero, since if the expected regret of a policy is zero, then it means that the policy knows the best action without playing all the other actions. Fortunately, the answer of the question is known for the classical multi-armed bandit problem, in other words, the tight lower bound of regret of strongly consistent policy is proved [12]. Strongly consistency demands policy to that work well for any situation. Detailed definition of strong consistency is in the main article.

In this thesis, we consider a property of policies called asymptotic optimality. If the expected regret of a given policy is equal to the lower bound of expected regret in the regime that the total number of plays goes to infinity, then the given policy is asymptotically optimal. As you can read, the asymptotic optimality is very strong condition. But, surprisingly, three of above-mentioned policies (KL-UCB, DMED, Thompson Sampling) satisfy such an optimality [7, 9, 11]. Consequently, these three policies are called asymptotically optimal policies.

In this thesis, we mainly do regret analysis for two generalized multi-armed bandit problems. First one is about a budgeted multi-armed bandit problem [13], which introduces a concept of cost in multi-armed bandit problem. Latter one is about a matching-selection bandit problem [14], a specific instance of the combinatorial bandit problem.

A budgeted multi-armed bandit problem is one kind of multi-armed bandit problem that incurs a cost when choosing an action and stops repetition if cumulative cost exceeds a given budget. This problem has several variants according to the generation model of cost. The earlier one assumes that the cost is constant that depends on just only the action [15, 16]. Later, a model that cost is generated according to action-aware distribution, has been proposed [13, 17]. we focused on the action-aware probabilistic model. For

this problem, we proposed the policy KL-UCB-SC (KL-UCB for Stochastic Costs) and proved its asymptotic optimality. This is proved by two steps. The first step is to prove the upper bound of expected regret of KL-UCB-SC. The second step is to prove the lower bound of expected regret over any strongly consistent policies for the budgeted multi-armed bandit problem.

Unfortunately, an efficient implementation of KL-UCB-SC is not known. Consequently, KL-UCB-SC requires so much computation time, which can be problematic. A prohibitive computation time for KL-UCB-SC came from the fact that the index to be computed for KL-UCB-SC is obtained by solving a complicated nonlinear optimization problem. For this problem, we further propose a policy UCB-SC (UCB for Stochastic Costs). UCB-SC is derived from KL-UCB-SC by the same procedure which is used for deriving the UCB from KL-UCB. We show that the index to be computed for UCB-SC can be written in a closed-form.

Therefore, UCB-SC can make decisions very rapidly compared to KL-UCB-SC. Of course, UCB-SC suffered a little too much expected regret as a compensation of computation speed.

In order to confirm the empirical characteristics of our proposed policies, we conduct a number of computational simulations. We report that the results of these simulations support the theoretical results. More precisely, an average regret by KL-UCB-SC+ was very small, UCB-SC+ worked as well as KL-UCB-SC+ in a kind of ideal situation, and UCB-SC+ runs with one thousandth computation time in comparison with KL-UCB-SC+. The plus-ed policies are the variants of the original policies that are made by applying a well-known tune-up technique [7, 9, 18].

A combinatorial bandit problem is the multi-armed bandit problem in which a player can choose a set of actions (which satisfies given condition) simultaneously [14, 19–22]. In this problem, a reward is generated for each chosen action and its total will be accumulated. The goal of a player is to maximize the cumulative reward.

There are two kinds of reward observation models for the combinatorial bandit problem. First one is called “full-bandit”. In this setting, a reward is observed as a scalar, that is, the sum of all the received rewards from chosen actions for the selected set of actions [22]. Second one is called “semi-bandit”. In this setting, rewards from chosen actions can be observed individually [14, 19–21]. This thesis focuses on semi-bandit setting.

Matching-selection bandit problem is a one specific class of combinatorial semi-bandit problem, where a set of action is written as an edge set of complete bipartite graph $K_{M,N}$ ($N \geq M \geq 1$) and each selection (subset of edge) must satisfy matching constraint.

On combinatorial (semi-)bandit problem, the difficulty of the regret anal-

ysis depends on the combinatorial structure. For instance, if a constraint belongs to the matroid (basis) constraint, a tight lower bound of expected regret was shown and an asymptotically optimal policy is already designed [23]. We remark that “multiple plays” setting [19, 24, 25], in which a set of m actions can be chosen for a given m , can be written as a matroid constraint. On the other hand, as for a matching constraint, computationally efficient policies are proposed [14, 20]. But a non-trivial lower bound of expected regret is still unknown, although such a constraint belongs to a special case of matroid intersection.

In this thesis, we analyzed a policy LLR [20] (Learning with Linear Rewards), which is a policy for general combinatorial semi-bandit problem in the case with matching-selection bandit problem by introducing a new technique. As a result, an upper bound of expected regret became smaller than known result for matching-selection bandit problem by a factor of $\Theta(M^{2/3})$.

This thesis is organized as follows.

We summarized our results in this chapter.

Chapter 2 reviews some fundamental results on the classical multi-armed bandit problem. This chapter contains the definition of the classical multi-armed bandit problem, definitions and regret upper bounds of several well-known policies, and the tight lower bound of regret over any strongly consistent policies.

In Chapter 3, we consider a budgeted multi-armed bandit problem. This chapter is mainly dedicated to show that a policy KL-UCB-SC is asymptotically optimal. Definition and derivation of UCB-SC, a computationally efficient variant of KL-UCB-SC is also included. At last of this chapter, results of computational experiments are reported.

In Chapter 4, we consider a matching-selection bandit problem. In this chapter, we show an upper bound of expected regret of LLR that is smaller than the known result in a matching-selection bandit problem.

Finally, Chapter 5 concludes this thesis.

Chapter 2

The Classical Multi-Armed Bandit Problem

2.1 Problem Formulation

2.1.1 Definition

The classical multi-armed bandits is a kind of one-player repeated stochastic game, described in Algorithm 1 as a form of pseudo-code. At each time t , a player chooses one element $I(t)$ from the given action set A , then receives a scalar reward $X_{I(t)}(t)$ according to the time and chosen action. This procedure is repeated until the time reaches fixed number of decision T .

In the classical formulation, an action set A is set to finite set. We will write the size of action set as $K = |A|$ and write each action as an integer 1 to K .

We assume that reward $X_i(t)$ have finite support for all $i \in A$ and $t = 1, 2, \dots$. Without loss of generality, the support is assumed to be the unit interval $[0, 1]$. Moreover, we assume that the rewards are stochastic. For each action $i \in A$, there is a probabilistic distribution ν_i on $[0, 1]$. Given this, when an action i is chosen at time t , reward is generated as $X_i(t) \sim \nu_i$. Any distinctive pair of reward are stochastically independent. We will write the expected reward of action i as $\mu_i = \mathbb{E}[X_i(t)]$. The *best action* $* \in A$ is defined as an action whose expected reward is maximum, that is,

$$* = \arg \max_{i \in A} \mu_i.$$

For simplicity, the best action is assumed to be unique.

As of choosing an action for time t , the player only can use past decision and according reward up to time $t - 1$ as a knowledge. Bringing the knowl-

edge, a rule for choosing action is called *policy*. More formally, let $\mathcal{H}(t)$ be the *history* of a game up to time t , defined as

$$\mathcal{H}(t) = \{ (I(s), X_{I(s)}(s)) \}_{s=1}^t.$$

Then a policy is defined as an infinite sequence of map¹ $\pi = \pi_1, \pi_2, \dots$, where $\pi_t: (A \times [0, 1])^{t-1} \rightarrow A$ is used to choose an action at time t . Namely,

$$I(t) = \pi_t(\mathcal{H}(t-1)).$$

In order to design a policy, we should settle the quantitative objective to be maximize or minimize. Actually, several objective have been considered on the same framework. Among them, the most conventional objective of the player is the maximizing of *cumulative reward* $G_\pi(T)$, defined as

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

Algorithm 1 The Classical Multi-Armed Bandits

Require: A : Action set.

- 1: **for** $t = 1, \dots, T$ **do**
 - 2: Choose $I(t) \in A$.
 - 3: Receive a reward $X_{I(t)}(t)$.
 - 4: **end for**
-

2.1.2 Cumulative Expected Pseudo-Regret

As described earlier, the grand objective of the policy is the maximizing cumulative reward. From the theoretical viewpoint, the quality of the policy is measured by the *cumulative expected pseudo-regret* of policy π , defined as

$$R_\pi(T) = \mathbb{E}[G_\pi(T) - G_{\pi^*}(T)].$$

where

$$\pi^* = \arg \max_{\pi \in \Pi} \mathbb{E}[G_\pi(T)]$$

and Π is the collection of any kinds of policy. We simply call cumulative expected pseudo-regret as regret hereafter.

¹In general, policy can be randomized. But we solely focus on deterministic policy throughout this paper.

This quantitative is the expected difference of the cumulative reward gained by the optimal policy π^* and that gained by a given policy π to be assessed. The optimal policy is the policy whose expected cumulative reward is maximal. On the classical setting, the optimal policy is the policy who chooses the best action for any time t . Namely,

$$\pi_t(\cdot) = I(t) = *.$$

The regret has a following simple alternative form

$$R_\pi(T) = \sum_{i: \Delta_i > 0} \Delta_i \mathbb{E}[N_i(T)] \quad (2.1)$$

where $\Delta_i = \mu_* - \mu_i$ and

$$N_i(t) = \#\{s \leq t \mid I(s) = i\}.$$

$N_i(t)$ is the count of the event of action i is chosen until the time t . Since the optimal policy always chooses the best action, choosing the non-best action is said to be a *mistake*. Hence, the quantity $N_i(t)$ for non-best action i is considered as a number of mistake caused by choosing an action i . In addition, Δ_i correspond to the regret per decision caused by choosing an action i . Thus, the regret is seemingly a weighed sum of expected number of mistake. Consequently, the essential part of regret is the expected number of mistake $\mathbb{E}[N_i(T)]$, as in Eq. (2.1).

2.2 Policies

2.2.1 The UCB (Upper Confidence Bound)

The UCB [3], originally called as UCB1 perhaps be the most famous policy for the classical multi-armed bandit problem.

Let

$$\hat{\mu}_{i,n} = \frac{1}{n} \sum_{s: N_i(s) \leq n, I(s)=i} X_{I(s)}(s).$$

be an estimator of μ_i using n samples and

$$\hat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{s \leq t, I(s)=i} X_{I(s)}(s).$$

be an estimator of μ_i after t th decision. Clearly, we have

$$\hat{\mu}_i(t) = \hat{\mu}_{i,N_i(t)}.$$

These notations are used interchangeably according to the context.

At time t , the UCB chooses an action $I(t)$ who maximizes some value $u_i(t)$.

$$I(t) = \arg \max_{i \in A} u_i(t) \quad (2.2)$$

The value $u_i(t)$ is sometimes called as action evaluation index. This framework is widely used for various policies. Therefore, a difference of policies are emerged in a definition of evaluation index $u_i(t)$.

Especially for the UCB, arm evaluation index $u_i(t)$ is defined as

$$u_i(t) = \hat{\mu}_i(t-1) + \sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}}, \quad (2.3)$$

where η is nonnegative real value. Here we employed an parametrized version in order to make later discussion easier. The detailed discussion of the η will be in later.

2.2.2 KL-UCB (Kullback-Leibler UCB)

Same as the UCB, KL-UCB uses Eq. (2.2) to determine the next action with different definition of evaluation index $u_i(t)$.

First we define a function $U(\mu, L; \mathcal{K})$ as

$$U(\mu, L; \mathcal{K}) = \max\{x \in [0, 1] \mid \mathcal{K}(\mu, x) \leq L\}.$$

for $\mu \in [0, 1]$, $L > 0$, and a function $\mathcal{K}: [0, 1] \times [0, 1] \rightarrow \mathbb{R}_*^+ (= \{x \mid x \geq 0\})$.

By using the function $U(\mu, L; \mathcal{K})$, action evaluation index for KL-UCB is defined as

$$u_i(t) = U\left(\hat{\mu}_i(t-1), \frac{\eta \ln(t)}{N_i(t-1)}, \mathcal{K}_{\mathcal{B}}\right) \quad (2.4)$$

where $\mathcal{K}_{\mathcal{B}}$ is the Kullback-Leibler divergence of Bernoulli distribution $\mathcal{B}(q)$ from Bernoulli distribution $\mathcal{B}(p)$, that is,

$$\mathcal{K}_{\mathcal{B}}(p, q) = p \ln \left(\frac{p}{q} \right) + (1-p) \ln \left(\frac{1-p}{1-q} \right).$$

This index can be computed by algorithms for root finding of nonlinear function, such as Newton-Raphson iteration.

Let us consider a valiant of KL-UCB that uses $\mathcal{K}_{\mathcal{N}}(p, q) = 2(q-p)^2$ for the argument of U , then we can immediately have

$$U(\mu, L; \mathcal{K}_{\mathcal{N}}) = \mu + \sqrt{\frac{L}{2}}. \quad (2.5)$$

By replacing the μ and L to $\hat{\mu}_i(t-1)$ and $\eta \ln(t)/N_i(t-1)$, as used in Eq. (2.4) respectively, it goes back to Eq. (2.3), the arm evaluation index of the UCB.

2.3 Regret Analysis of UCB Family

2.3.1 Known Facts

Here we just list several facts used for proving upper/lower bound of expected regret.

Fact 1 (Hoeffding's inequality [26]). *Let $X_1, \dots, X_n$ be the i.i.d. random variables taking values in $[0, 1]$ having common mean μ . Let*

$$\hat{\mu}_n = \frac{1}{n} \sum_{s=1}^n X_s.$$

Then, for any $a \geq 0$,

$$\Pr\{\hat{\mu}_n \geq \mu + a\} \leq \exp(-2na^2)$$

and

$$\Pr\{\hat{\mu}_n \leq \mu - a\} \leq \exp(-2na^2).$$

Fact 2 (Chernoff-Cramer's inequality). *Let $X_1, \dots, X_n$ be the i.i.d. random variable taking values in $[0, 1]$ having common mean μ . Let*

$$\hat{\mu}_n = \frac{1}{n} \sum_{s=1}^n X_s.$$

Then for any $\epsilon \geq 0$,

$$\Pr\{\mathcal{K}_{\mathcal{B}}(\mu, \hat{\mu}_n) \geq \epsilon\} \leq \exp(-n\epsilon).$$

Fact 3 (Pinsker's inequality). *For all $p, q \in [0, 1]$, the following inequality holds:*

$$\mathcal{K}_{\mathcal{B}}(p, q) \geq \mathcal{K}_{\mathcal{N}}(p, q).$$

Fact 4 (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 5 (Maximal law of large numbers [27]). *Let $X_1, \dots$ 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 $n \rightarrow \infty$, then

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

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

2.3.2 Regret Analysis of the UCB

As in Eq. (2.1), an essential part of expected regret is the $\mathbb{E}[N_i(T)]$, an expected number of mistake up to time T for all $i \neq *$. Following theorem shows a upper bound of that when a parameter η was chosen conservatively.

Theorem 1. *On the classical multi-armed bandit problem, if the action is chosen according to the UCB with $\eta > 2$,*

$$\mathbb{E}[N_i(T)] \leq \frac{2\eta \ln(T)}{\Delta_i^2} + \frac{\eta}{\eta - 2}.$$

*holds for any action $i \neq *$.*

Proof. An event $I(t) = i \neq *$ can be decomposed into three events.

$$\begin{aligned} \{I(t) = i\} &\subseteq \left\{ \hat{\mu}_i(t-1) - \sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} > \mu_i \right\} \\ &\cup \left\{ \hat{\mu}_*(t-1) + \sqrt{\frac{\eta \ln(t)}{2N_*(t-1)}} < \mu_* \right\} \\ &\cup \left\{ \begin{array}{l} \hat{\mu}_i(t-1) < \mu_i + \sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} \\ u_i(t) > u_*(t) : \\ \hat{\mu}_*(t-1) > \mu_* - \sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} \end{array} \right\} \end{aligned}$$

The probability of the first event can be bounded as

$$\begin{aligned}
 & \sum_{t=1}^T \Pr \left\{ \hat{\mu}_i(t-1) - \sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} > \mu_i \right\} \\
 &= \sum_{t=1}^T \sum_{n=1}^{t-1} \Pr \left\{ \hat{\mu}_{i,n} > \mu_i + \sqrt{\frac{\eta \ln(t)}{2n}}, N_i(t-1) = n \right\} \\
 &\leq \sum_{t=1}^T \sum_{n=1}^{t-1} \Pr \left\{ \hat{\mu}_{i,n} > \mu_i + \sqrt{\frac{\eta \ln(t)}{2n}} \right\}.
 \end{aligned}$$

By Hoeffding's inequality (Fact 1), we find that

$$\Pr \left\{ \hat{\mu}_{i,n} > \mu_i + \sqrt{\frac{\eta \ln(t)}{2n}} \right\} \leq t^{-\eta}.$$

Hence,

$$\begin{aligned}
 \sum_{t=1}^T \sum_{n=1}^{t-1} \Pr \left\{ \hat{\mu}_{i,n} > \mu_i + \sqrt{\frac{\eta \ln(t)}{2n}} \right\} &\leq \sum_{n=1}^T t^{-\eta} \\
 &\leq \sum_{t=1}^T t^{1-\eta}. \tag{2.6}
 \end{aligned}$$

Assume that $\eta \geq 2$, then by indefinite integral, Eq. (2.6) can be upper bounded as

$$\begin{aligned}
 \sum_{t=1}^T t^{1-\eta} &\leq \sum_{t=1}^{\infty} t^{1-\eta} \\
 &\leq \int_0^{\infty} t^{1-\eta} dt \\
 &= -\frac{1}{2-\eta} = \frac{1}{\eta-2}.
 \end{aligned}$$

In summary, if $\eta > 2$,

$$\sum_{t=1}^T \Pr \left\{ \hat{\mu}_i(t-1) - \sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} > \mu_i \right\} \leq \frac{1}{\eta-2}. \tag{2.7}$$

The upper bound for the second event can be bounded as

$$\sum_{t=1}^T \Pr \left\{ \hat{\mu}_*(t-1) + \sqrt{\frac{\eta \ln(t)}{2N_*(t-1)}} < \mu_* \right\} \leq \frac{1}{\eta-2}. \tag{2.8}$$

by the same procedure for the first event.

The third event can be simplified as

$$\left\{ \Delta_i \geq 2\sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} \right\}.$$

Hence, for any integer $Z_i \geq 1$,

$$\begin{aligned} & \sum_{t=1}^T \mathbb{I} \left\{ \Delta_i > 2\sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} \right\} \\ &= \sum_{t=1}^T \sum_{n=1}^{t-1} \mathbb{I} \left\{ \Delta_i > 2\sqrt{\frac{\eta \ln(t)}{2n}}, N_i(t-1) = n \right\} \\ &\leq \sum_{n=1}^{T-1} \sum_{t=n+1}^T \mathbb{I} \left\{ \Delta_i > 2\sqrt{\frac{\eta \ln(T)}{2n}} \right\} \cdot \mathbb{I}\{N_i(t-1) = n\} \\ &\leq Z_i + \sum_{n=Z_i}^{T-1} \mathbb{I} \left\{ \Delta_i > 2\sqrt{\frac{\eta \ln(T)}{2n}} \right\} \sum_{t=n+1}^T \mathbb{I}\{N_i(t-1) = n\} \end{aligned}$$

The second term goes zero when

$$Z_i \geq \frac{2\eta \ln(T)}{\Delta_i^2},$$

since

$$\begin{aligned} 2\sqrt{\frac{\eta \ln(T)}{2Z_i}} &\geq 2\sqrt{\frac{\eta \ln(T)}{2} \frac{\Delta_i^2}{2\eta \ln(T)}} \\ &= \Delta_i. \end{aligned}$$

In summary we have,

$$\begin{aligned} \sum_{t=1}^T \mathbb{I} \left\{ \Delta_i > 2\sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} \right\} &\leq \left\lceil \frac{2\eta \ln(T)}{\Delta_i^2} \right\rceil \\ &\leq \frac{2\eta \ln(T)}{\Delta_i^2} + 1 \end{aligned} \tag{2.9}$$

Finally, putting Eq. (2.7), Eq. (2.8), and Eq. (2.9) together, we get

$$\begin{aligned}
\mathbb{E}[N_i(T)] &= \mathbb{E} \left[\sum_{t=1}^T \mathbb{I}\{I(t) = i\} \right] \\
&\leq \mathbb{E} \left[\sum_{t=1}^T \mathbb{I} \left\{ \Delta_i > 2\sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} \right\} \right] \\
&\quad + \sum_{t=1}^T \Pr \left\{ \hat{\mu}_i(t-1) - \sqrt{\frac{\eta \ln(t)}{2N_i(t-1)}} > \mu_i \right\} \\
&\quad + \sum_{t=1}^T \Pr \left\{ \hat{\mu}_*(t-1) + \sqrt{\frac{\eta \ln(t)}{2N_*(t-1)}} \leq \mu_* \right\} \\
&\leq \frac{2\eta \ln(T)}{\Delta_i^2} + 1 + \frac{2}{\eta - 2} \\
&= \frac{2\eta \ln(T)}{\Delta_i^2} + \frac{\eta}{\eta - 2}
\end{aligned}$$

if $\eta > 2$. □

With a help of Eq. (2.1) and Theorem 1, we immediately have an upper bound of expected regret for the UCB.

Corollary 1. *Let π be the UCB with $\eta > 2$. On the classical multi-armed bandit problem,*

$$R_\pi(T) \leq \sum_{i: \Delta_i > 0} \left(\frac{2\eta \ln(T)}{\Delta_i} + \frac{2\eta \Delta_i}{\eta - 2} \right).$$

holds.

As shown in later, this analysis is not tight. But, due to its simplicity, this kind of proof technique is still useful if one only curious whether the expected regret is bounded by $O(\ln(T))$ as a function of T .

2.3.3 Regret Analysis of KL-UCB

We will show an upper bound of expected regret for KL-UCB.

Theorem 2. *On the classical multi-armed bandit problem, if the action is chosen according to KL-UCB with $\eta = 1$,*

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E}[N_i(T)]}{\ln(T)} = \frac{1}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu_*)},$$

hence,

$$\mathbb{E}[N_i(T)] \leq \frac{\ln(T)}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu_*)} + o(\ln(T))$$

holds for any action $i \neq *$.

Proof. See [8]. □

Same as the UCB, Theorem 2 immediately derives an upper bound of expected regret for KL-UCB.

Corollary 2. *Let π be KL-UCB with $\eta = 1$. On the classical multi-armed bandit problem,*

$$R_\pi(T) \leq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu_*)} \ln(T) + o(\ln(T))$$

holds.

The same proof can be applied for the UCB with in case of $\eta = 1$,

Corollary 3. *Let π be the UCB with $\eta = 1$. On the classical multi-armed bandit problem, then*

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E}[N_i(T)]}{\ln(T)} = \frac{1}{\mathcal{K}_{\mathcal{N}}(\mu_i, \mu_*)} = \frac{1}{2\Delta_i^2},$$

holds for any action $i \neq *$. Hence,

$$R_\pi(T) \leq \sum_{i: \Delta_i > 0} \frac{\ln(T)}{2\Delta_i^2} + o(\ln(T))$$

holds.

In comparison of Corollary 1 and Corollary 3, later achieved asymptotically smaller upper bound by a factor of 8.

2.4 Lower Bound of Regret

When we discuss about the lower bound of expected regret, we should focus only on the policies that works fine for any situation. Such a concept is formulated as “strong consistency”, defined in the following.

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

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E}[N_i(T)]}{T^p} = 0$$

holds for any distributions.

We will write a collection of strong consistent policies as $\Pi^* \subset \Pi$.

Under a strongly consistent condition, a lower bound of expected regret were proved.

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

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E}[N_i(T)]}{\ln(T)} = \frac{1}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu_*)}.$$

In other words,

$$\mathbb{E}[N_i(T)] = (1 - o(1)) \frac{\ln(T)}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu_*)}.$$

Proof. See [12] or [27]. □

Corollary 4. *Let the rewards be drawn from Bernoulli distributions with parameters $\{\mu_i\}_{i \in A}$. Then the regret is lower-bounded as:*

$$\lim_{T \rightarrow \infty} \frac{\inf\{R_\pi(T) \mid \pi \in \Pi^*\}}{\ln(T)} = \sum_{i: \Delta_i > 0} \frac{\Delta_i}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu_*)}.$$

In other words,

$$\inf\{R_\pi(T) \mid \pi \in \Pi^*\} = \sum_{i: \Delta_i > 0} \frac{\Delta_i}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu_*)} \ln(T) - o(\ln(T)).$$

Compared to the bounds proved in Theorem 2 and Theorem 3 (or correspond corollaries), following optimality of KL-UCB can be shown.

Corollary 5. *On the classical multi-armed bandit problem, if all reward distributions are Bernoulli, then KL-UCB is asymptotically optimal. That is, the coefficient of the logarithmic factor of regret as a function of total number of plays T cannot be improved. More formally, the following equation is true if π is KL-UCB.*

$$\lim_{T \rightarrow \infty} \frac{R_\pi(T)}{\inf\{R_{\pi'}(T) \mid \pi' \in \Pi^*\}} = 1$$

Chapter 3

Budgeted Multi-Armed Bandit Problem

3.1 Problem Formulation

3.1.1 Definition

The budgeted multi-armed bandits [13] is described as an one-player game in which a player repeatedly chooses one of actions among given action set A until he/she uses up a budget. At each round $t = 1, 2, \dots$, a player chooses an action $I(t) \in A$, and then he/she pays a cost $Y_{I(t)}(t) \in [0, 1]$ and receives a reward $X_{I(t)}(t) \in [0, 1]$. The rewards $X_i(1), X_i(2), \dots$ and costs $Y_i(1), Y_i(2), \dots$ 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 $Y_i(t)$ of unchosen actions $i \neq I(t)$ are not revealed to the player. For analytical reasons, we assume that $\mu_i, \tau_i \in (0, 1)$. Let $\tau_{\min}$ and $\tau_{\max}$ denote $\min\{\tau_i \mid i \in A\}$ and $\max\{\tau_i \mid i \in A\}$, respectively. Let

$$C_t = \sum_{s=1}^t Y_{I(s)}(s)$$

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 2.

Let us study the player's policy $\pi = \pi_1, \pi_2, \dots$ for choosing action $I(t) = \pi_t(\mathcal{H}(t-1))$. For budgeted bandits, the history would be represented as a sequence of 3-tuple $\mathcal{H}(t) = \{(I(s), X_{I(s)}(s), Y_{I(s)}(s))\}_{s=1}^t$ and each map for time t should be $\pi_t: (A \times [0, 1] \times [0, 1])^{t-1} \rightarrow A$. We evaluate policy π by

Algorithm 2 Budgeted Multi-Armed Bandits

```

1:  $t \leftarrow 1$ 
2:  $C_0 \leftarrow 0$ 
3: while  $C_{t-1} \leq B$  do
4:   Choose an action  $I(t) \in A$ .
5:   A cost  $Y_{I(t)}(t)$  is revealed.
6:    $C_t \leftarrow C_{t-1} + Y_{I(t)}(t)$ 
7:   if  $C_t \leq B$  then
8:     Receive reward  $X_{I(t)}(t)$ .
9:      $t \leftarrow t + 1$ .
10:  end if
11: end while
    
```

using *regret*, which is defined in the following. Let

$$T(B) = \max \left\{ t : \sum_{s=1}^t Y_{I(s)}(s) \leq B \right\}$$

be the final round when the player could gain a reward. The cumulative reward of the player who follows policy π is

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

Let

$$\pi^* = \arg \max_{\pi \in \Pi} \mathbb{E}[G_\pi(B)]$$

be the best policy where Π denotes any collection of the policy. Then, the performance of a policy π is measured by the *pseudo regret*

$$R_\pi(B) = \mathbb{E}[G_{\pi^*}(B)] - \mathbb{E}[G_\pi(B)].$$

3.1.2 Considering a Stopping Time on Regret

Unlike to the classical problem, an exact evaluation of $R_\pi(B)$ 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 $*$ $\in A$ is defined as

$$* = \arg \max_{i \in A} \frac{\mu_i}{\tau_i}.$$

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

$$N_i(t) = \sum_{s=1}^t \mathbb{I}\{I(s) = i\}$$

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 regret¹. Thus we again define this amount as the regret.

Theorem 4 (Lemma 2 by [17]). *For any policy π of the budgeted multi-armed bandits with budget B , we have*

$$\left| R_\pi(B) - \tilde{R}_\pi(B) \right| \leq 2 \frac{\mu_*}{\tau_*}. \quad (3.1)$$

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

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 1. *For any policy π , we have*

$$\tilde{R}_\pi(B) \leq \sum_{i: \Delta_i > 0} \tau_i \Delta_i \left(\mathbb{E} \left[N_i \left(\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor \right) \right] + \frac{K^2}{2\tau_{\min}^2} \right) + o(1) \quad (3.2)$$

and

$$\tilde{R}_\pi(B) \geq \sum_{i: \Delta_i > 0} \tau_i \Delta_i \mathbb{E} \left[N_i \left(\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor + 1 \right) \right] - o(1), \quad (3.3)$$

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

¹Note that, Lemma 2 by Xia et al. [17] 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.

Proof. Let $Y_{i,s}$ denote the cost $Y_i(t)$ for the player's s -th choice of action i . We first derive Ineq (3.2). For $T = 1, 2, \dots$, let $z_T = \tau_{\min} T / (KB)$. Then we have

$$\begin{aligned} \Pr\{T(B) \geq T\} &= \Pr\left\{T(B) \geq \frac{z_T KB}{\tau_{\min}}\right\} \\ &\leq \Pr\left\{\bigcup_{i \in [K]} \left\{\sum_{s=1}^{\left\lceil \frac{z_T B}{\tau_{\min}} \right\rceil} Y_{i,s} \leq B\right\}\right\} \\ &\leq \sum_{i \in A} \Pr\left\{\sum_{s=1}^{\left\lceil \frac{z_T B}{\tau_{\min}} \right\rceil} Y_{i,s} \leq B\right\} \\ &\leq \sum_{i \in A} \Pr\left\{\frac{1}{\left\lceil \frac{z_T B}{\tau_{\min}} \right\rceil} \sum_{s=1}^{\left\lceil \frac{z_T B}{\tau_{\min}} \right\rceil} Y_{i,s} \leq \tau_i - \frac{z_T - 1}{z_T} \tau_{\min}\right\} \end{aligned}$$

since

$$\frac{B}{\left\lceil \frac{z_T B}{\tau_{\min}} \right\rceil} \leq \frac{\tau_{\min}}{z_T} \leq \frac{\tau_{\min}}{z_T} + \tau_i - \tau_{\min} = \tau_i - \frac{z_T - 1}{z_T} \tau_{\min}.$$

Then by Hoeffding's inequality (Fact 1), we find that

$$\begin{aligned} &\Pr\left\{\frac{1}{\left\lceil \frac{z_T B}{\tau_{\min}} \right\rceil} \sum_{s=1}^{\left\lceil \frac{z_T B}{\tau_{\min}} \right\rceil} Y_{i,s} \leq \tau_i - \frac{z_T - 1}{z_T} \tau_{\min}\right\} \\ &\leq \exp\left(-2 \left(\frac{z_T B}{\tau_{\min}}\right) \left(\frac{z_T - 1}{z_T}\right)^2 \tau_{\min}^2\right) \\ &= \exp\left(-2 \left(\frac{(z_T - 1)^2}{z_T}\right) \tau_{\min} B\right). \end{aligned}$$

Consequently,

$$\begin{aligned} \Pr\{T(B) \geq T\} &\leq \sum_{i \in A} \exp\left(-2 \left(\frac{(z_T - 1)^2}{z_T}\right) \tau_{\min} B\right) \\ &= K \exp\left(-2 \left(\frac{(z_T - 1)^2}{z_T}\right) \tau_{\min} B\right) \end{aligned} \tag{3.4}$$

$$\leq K \exp\left(-2 \tau_{\min} B \left(\frac{\tau_{\min} T}{KB} - 2\right)\right). \tag{3.5}$$

For any action i , we have

$$\begin{aligned}\mathbb{E}[N_i(T(B))] &= \sum_{T=1}^{\infty} \mathbb{E}[N_i(T)] \Pr\{T(B) = T\} \\ &\leq \sum_{T=1}^{\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor} \mathbb{E}[N_i(T)] \Pr\{T(B) = T\} + \sum_{\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1}^{\infty} T \cdot \Pr\{T(B) = T\}\end{aligned}$$

since $N_i(T) \leq T$. The first term can be bounded as

$$\begin{aligned}\sum_{T=1}^{\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor} \mathbb{E}[N_i(T)] \Pr\{T(B) = T\} &\leq \sum_{T=1}^{\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor} \mathbb{E} \left[N_i \left(\frac{2KB}{\tau_{\min}} \right) \right] \Pr\{T(B) = T\} \\ &= \mathbb{E} \left[N_i \left(\frac{2KB}{\tau_{\min}} \right) \right] \sum_{T=1}^{\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor} \Pr\{T(B) = T\} \\ &\leq \mathbb{E} \left[N_i \left(\frac{2KB}{\tau_{\min}} \right) \right].\end{aligned}\tag{3.6}$$

The second term will also be decomposed into two terms.

$$\begin{aligned}&\sum_{T=\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1}^{\infty} T \cdot \Pr\{T(B) = T\} \\ &= \sum_{T=\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1}^{\infty} T \cdot \Pr\{T(B) \geq T\} - T \cdot \Pr\{T(B) \geq T + 1\} \\ &= \left(\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1 \right) \Pr \left\{ T(B) \geq \left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1 \right\} + \sum_{T=\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 2}^{\infty} \Pr\{T(B) \geq T\}\end{aligned}$$

Now we can get

$$\Pr \left\{ T(B) \geq \left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1 \right\} \leq K e^{-\tau_{\min} B}$$

using Eq. (3.4) and following fact

$$z_{\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1} = \frac{\tau_{\min} \left(\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1 \right)}{KB} \geq 2.$$

Consequently,

$$\left(\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1\right) \Pr \left\{ T(B) \geq \left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 1 \right\} = o(1) \quad (3.7)$$

as a function of B . From Eq. (3.5), last term will be bounded as

$$\begin{aligned} \sum_{T=\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 2}^{\infty} \Pr\{T(B) \geq T\} &\leq \sum_{T=\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 2}^{\infty} K \exp \left(-2\tau_{\min} B \left(\frac{\tau_{\min} T}{KB} - 2 \right) \right) \\ &= K \sum_{T=\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 2}^{\infty} \exp \left(4\tau_{\min} B - \frac{2\tau_{\min}^2 T}{K} \right) \\ &= K e^{4\tau_{\min} B} \cdot \frac{\exp \left(-\frac{2\tau_{\min}^2}{K} \left(\left\lfloor \frac{2KB}{\tau_{\min}} \right\rfloor + 2 \right) \right)}{1 - e^{-2\tau_{\min}^2/K}} \\ &\leq K e^{4\tau_{\min} B} \cdot \frac{\exp \left(-\frac{2\tau_{\min}^2}{K} \left(\frac{2KB}{\tau_{\min}} + 1 \right) \right)}{1 - e^{-2\tau_{\min}^2/K}} \\ &= K e^{4\tau_{\min} B} \cdot \frac{\exp \left(-4\tau_{\min} B - \frac{2\tau_{\min}^2}{K} \right)}{1 - e^{-2\tau_{\min}^2/K}} \\ &= K \cdot \frac{e^{-2\tau_{\min}^2/K}}{1 - e^{-2\tau_{\min}^2/K}}. \end{aligned} \quad (3.8)$$

Up to here, Eq. (3.8) is bounded by a constant independent to budget B . Moreover, its simpler upper bound can be shown as

$$K \cdot \frac{e^{-2\tau_{\min}^2/K}}{1 - e^{-2\tau_{\min}^2/K}} \leq \frac{K^2}{2\tau_{\min}^2},$$

since $e^{-x}/(1 - e^{-x}) = 1/(e^x - 1) \leq 1/x$ for any $x \geq 0$.

Putting Eq. (3.6), Eq. (3.7), and Eq. (3.8) together, we find

$$\mathbb{E}[N_i(T(B))] \leq \mathbb{E} \left[N_i \left(\frac{2KB}{\tau_{\min}} \right) \right] + \frac{K^2}{2\tau_{\min}^2} + o(1).$$

Consequently, we find

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

which proof the Eq. (3.2).

Next we derive Ineq. (3.3). First, we have

$$\begin{aligned}
 \Pr \left\{ T(B) \leq \left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor \right\} &\leq \Pr \left\{ \bigcup_{i \in A} \left\{ \sum_{s=1}^{\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor} Y_{i,s} \geq \frac{B}{K} \right\} \right\} \\
 &\leq \sum_{i \in A} \Pr \left\{ \sum_{s=1}^{\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor} Y_{i,s} \geq \frac{B}{K} \right\} \\
 &\leq \sum_{i \in A} \Pr \left\{ \frac{1}{\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor} \sum_{s=1}^{\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor} Y_{i,s} \geq \tau_i + \tau_{\max} \right\}
 \end{aligned}$$

since

$$\frac{B}{K} \frac{1}{\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor} \geq 2\tau_{\max} \geq \tau_i + \tau_{\max}.$$

Then by Hoeffding's inequality (Fact 1), we find that

$$\Pr \left\{ \frac{1}{\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor} \sum_{s=1}^{\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor} Y_{i,s} \geq \tau_i + \tau_{\max} \right\} \leq \exp \left(-2 \left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor \tau_{\max}^2 \right).$$

Consequently,

$$\begin{aligned}
 \Pr \left\{ T(B) \leq \left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor \right\} &\leq \sum_{i \in A} \exp \left(-2 \left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor \tau_{\max}^2 \right) \\
 &\leq K \exp \left(-2 \left(\frac{B}{2K\tau_{\max}} + 1 \right) \tau_{\max}^2 \right) \quad (3.9)
 \end{aligned}$$

For any action i ,

$$\begin{aligned}
 & \mathbb{E}[N_i(T(B))] \\
 &= \sum_{T=1}^{\infty} \mathbb{E}[N_i(T)] \Pr\{T(B) = T\} \\
 &\geq \sum_{T=\lfloor \frac{B}{2K\tau_{\max}} \rfloor + 1}^{\infty} \mathbb{E}[N_i(T)] \Pr\{T(B) = T\} \\
 &\geq \mathbb{E} \left[N_i \left(\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor + 1 \right) \right] \left(1 - \Pr \left\{ T(B) \leq \left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor \right\} \right) \\
 &\geq \mathbb{E} \left[N_i \left(\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor + 1 \right) \right] \\
 &\quad + \mathbb{E} \left[N_i \left(\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor + 1 \right) \right] \Pr \left\{ T(B) \leq \left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor \right\}
 \end{aligned}$$

By using Eq (3.9) and a fact $N_i(T) \leq T$ for any $T \geq 1$, the second term can be bounded as

$$\begin{aligned}
 & \mathbb{E} \left[N_i \left(\left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor + 1 \right) \right] \Pr \left\{ T(B) \leq \left\lfloor \frac{B}{2K\tau_{\max}} \right\rfloor \right\} \\
 &\leq \left(\frac{B}{2K\tau_{\max}} + 1 \right) K \exp \left(-2 \left(\frac{B}{2K\tau_{\max}} + 1 \right) \tau_{\max}^2 \right) \\
 &= \left(\frac{B}{2\tau_{\max}} + K \right) \exp \left(-2 \left(\frac{B}{2K\tau_{\max}} + 1 \right) \tau_{\max}^2 \right) \\
 &= o(1),
 \end{aligned}$$

as a function of B .

Then we find that

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

□

With the help of Lemma 1, we can derive the upper/lower bound of regret as a function of the budget B if we succeeded to derive the upper/lower bound

of $\mathbb{E}[N_i(T)]$ for fixed T . Compared with the analysis by Xia et al. [17], that always takes the budget into consideration, Lemma 1 makes the analysis much easier.

3.2 Related Works

As for the budgeted multi-armed bandits, KUBE [16], the 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 (fractional-) KUBE does not solve unbounded knapsack problems with estimated mean rewards, the strong consistency is proved. [13] extended the problem so as to allow the action costs to be distributed.

UCB-BV family [13] and PD-BwK [28] are policies of the naïve “two-bound approach” for this problem. BTS [17], 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.

3.3 Policies

3.3.1 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,

$$I(t) = \arg \max_{i \in A} u_i(t) \quad (3.10)$$

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 the empirical mean reward and cost up to n samples be

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

and

$$\hat{\tau}_{i,n} = \frac{1}{n} \sum_{s: N_i(s) \leq n, I(s)=i} Y_i(s)$$

respectively. We also define the empirical mean reward and cost up to time t as

$$\hat{\mu}_i(t) = \hat{\mu}_{i,N_i(t)} = \frac{1}{N_i(t)} \sum_{s \leq t, I(s)=i} X_i(s)$$

and

$$\hat{\tau}_i(t) = \hat{\tau}_{i,N_i(t)} = \frac{1}{N_i(t)} \sum_{s \leq t, I(s)=i} Y_i(s).$$

Then the index $u_i(t)$ will be calculated as

$$u_i(t) = U \left(\hat{\mu}_i(t-1), \hat{\tau}_i(t-1), \frac{\ln(t)}{N_i(t)}, \mathcal{K}_{\mathcal{B}} \right),$$

where $U(\mu, \tau, D, \mathcal{K})$ is defined by following nonlinear optimization problem

$$U(\mu, \tau, L, \mathcal{K}) = \max \left\{ \frac{x}{y} : (x, y) \in [0, 1]^2, \mathcal{K}(\mu, x) + \mathcal{K}(\tau, y) \leq L \right\}.$$

Intuitively, the feasible solution $\{(x, y) \mid \mathcal{K}(\mu, x) + \mathcal{K}(\tau, y) \leq L\}$ represents the two dimensional region where (x, y) is within L from (μ, τ) in terms of the pseudo-distance $d((\mu, \tau), (x, y)) = \mathcal{K}(\mu, x) + \mathcal{K}(\tau, y)$.

The advantages of the index $u_i(t)$ are that (i) the confidence bound derived from the KL divergence using Cramer-Chernoff inequality (Fact 2) is tighter than the one derived from quadratic divergence using Hoeffding's inequality (Fact 1), as is understood from Pinsker's inequality (Fact 3). 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.

3.3.2 UCB-SC

As described in Section 2.2.1, The UCB for classical multi-armed bandit problem can be derived by replacing the divergence function of KL-UCB. A policy UCB-SC can be defined as exactly the same way, that is,

$$I(t) = \arg \max_{i \in \mathcal{A}} U \left(\hat{\mu}_i(t-1), \hat{\tau}_i(t-1), \frac{\ln(t)}{N_i(t-1)}; \mathcal{K}_{\mathcal{N}} \right).$$

In comparison of the UCB and KL-UCB, an advantage of the UCB is fast computation time. It is due to an action evaluation index for the UCB

has a closed-form. Hence, we expected that an action evaluation index of UCB-SC has a closed-form. Following theorem shows a closed-form of an action evaluation index for UCB-SC.

Theorem 5.

$$U(\mu, \tau, L; \mathcal{K}_N) = \begin{cases} \frac{\mu + \frac{\tau}{\sqrt{\mu^2 + \tau^2 - L/2}} \sqrt{L/2}}{\tau - \frac{\mu}{\sqrt{\mu^2 + \tau^2 - L/2}} \sqrt{L/2}} & (\tau > L/2) \\ \infty & (\text{otherwise}) \end{cases} \quad (3.11)$$

Proof. First we consider the case of $\tau > \sqrt{L/2}$. From a viewpoint of geometry, $U(\mu, \tau, L; \mathcal{K}_N)$ is the maximum gradient u among those of the two lines $ux - y = 0$ whose Euclidean distance from (τ, μ) is exactly $\sqrt{L/2}$. Hence, by using a formula of the distance between a point and a line, $U(\mu, \tau, L; \mathcal{K}_N)$ is the larger solution u of the following equation

$$\frac{|u\tau - \mu|}{\sqrt{u^2 + 1}} = \sqrt{\frac{\delta}{2}},$$

which is solved by quadratic formula as

$$u = \frac{\mu\tau + \sqrt{(L/2)(\mu^2 + \tau^2 - L/2)}}{\tau^2 - L/2}. \quad (3.12)$$

By multiplying

$$\frac{\tau - \frac{\mu}{\sqrt{\mu^2 + \tau^2 - L/2}} \sqrt{L/2}}{\tau^2 - L/2}$$

to the denominator of Eq. (3.12), it will immediately be the same as a de-

nominator of Eq. (3.11). The numerator will be

$$\begin{aligned}
 & \left(\mu\tau + \sqrt{(L/2)(\mu^2 + \tau^2 - L/2)} \right) \left(\frac{\tau - \frac{\mu}{\sqrt{\mu^2 + \tau^2 - L/2}} \sqrt{L/2}}{\tau^2 - L/2} \right) \\
 &= \frac{\mu\tau^2 - \frac{\mu^2\tau\sqrt{L/2}}{\sqrt{\mu^2 + \tau^2 - L/2}} + \tau\sqrt{(L/2)(\mu^2 + \tau^2 - L/2)} - \mu(L/2)}{\tau^2 - L/2} \\
 &= \frac{\mu\tau^2 - \mu(L/2)}{\tau^2 - L/2} + \frac{\tau\sqrt{(L/2)(\mu^2 + \tau^2 - L/2)} - \frac{\mu^2\tau\sqrt{L/2}}{\sqrt{\mu^2 + \tau^2 - L/2}}}{\tau^2 - L/2} \\
 &= \mu + \frac{\tau\sqrt{L/2} \left(\sqrt{\mu^2 + \tau^2 - L/2} - \frac{\mu^2}{\sqrt{\mu^2 + \tau^2 - L/2}} \right)}{\tau^2 - L/2} \\
 &= \mu + \frac{\tau\sqrt{L/2} \left(\frac{\mu^2 + \tau^2 - L/2}{\sqrt{\mu^2 + \tau^2 - L/2}} - \frac{\mu^2}{\sqrt{\mu^2 + \tau^2 - L/2}} \right)}{\tau^2 - L/2} \\
 &= \mu + \frac{\tau\sqrt{L/2} \frac{\tau^2 - L/2}{\sqrt{\mu^2 + \tau^2 - L/2}}}{\tau^2 - L/2} \\
 &= \mu + \frac{\tau}{\sqrt{\mu^2 + \tau^2 - L/2}} \sqrt{L/2},
 \end{aligned}$$

which is the same as a numerator of Eq. (3.11).

In case of $\tau \leq \sqrt{L/2}$, the objective function can take arbitrarily large value since a feasible solution of $U(\mu, \tau, L; \mathcal{K}_{\mathcal{N}})$ contains a point $(\mu, 0)$. $\square$

Although Eq. (3.11) is available for a closed-form action evaluation index for UCB-SC, we derived Eq. (3.11), since an action evaluation index for the UCB can be expressed as $\mu + \sqrt{L/2}$ (Eq. (2.5)). Note that the denominator and numerator of Eq. (3.11) do not correspond to x and y of $U(\mu, \tau, L; \mathcal{K}_{\mathcal{N}})$ respectively.

3.4 Analytic Results

3.4.1 Dissimilarity of an Action between the Best Action

This section is devoted to show the upper bound of regret when a player obeys KL-UCB-SC. Let

$$D_i = \min_{a > 0} (\mathcal{K}_{\mathcal{B}}(\mu_i, a\mu_*) + \mathcal{K}_{\mathcal{B}}(\tau_i, a\tau_*)).$$

This quantity characterizes how different the best action $*$ from the given action i .

In the context of standard multi-armed bandit problem, Lai and Robbins [12] showed that a strongly consistent algorithm must choose action i until the number of draws $N_i(t)$ satisfies $\ln(t) \simeq N_i(t) \mathcal{K}_{\mathcal{B}}(\hat{\mu}_i, \mu_*)$. Note that the quantity $N_i(t) \mathcal{K}_{\mathcal{B}}(\hat{\mu}_i, \mu_*)$ 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 $\ln(t) \simeq N_i(t) \min(\mathcal{K}_{\mathcal{B}}(\hat{\mu}_i, a\mu_*) + \mathcal{K}_{\mathcal{B}}(\hat{\tau}_i, a\tau_*))$ 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.

3.4.2 Upper Bound of the Expected Regret for KL-UCB-SC

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 rounds for any $i \neq *$ 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 3 and 5 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 $\mu_*(\epsilon)/\tau_*(\epsilon) (< \mu_*/\tau_*)$.

Precede to show the main statement, we prepare a number of propositions and lemmas.

Proposition 1 ([9]). *Let $f(\mu, \mu_1) = (\mu - \mu_1)^2 / (2\mu(1 - \mu_1))$. Then,*

$$\mathcal{K}_{\mathcal{B}}(\mu_2, \mu) - \mathcal{K}_{\mathcal{B}}(\mu_2, \mu_1) \geq f(\mu, \mu_1) > 0 \quad (3.13)$$

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

$$\mathcal{K}_{\mathcal{B}}(\tau_2, \tau) - \mathcal{K}_{\mathcal{B}}(\tau_2, \tau_1) \geq f(\tau_1, \tau) > 0 \quad (3.14)$$

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

Proof. The proof of Eq. (3.13) is in the original paper [9]. Eq. (3.14) can be proved in the same way. $\square$

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

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

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

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

$\square$

Lemma 2. *Let $\epsilon > 0$ be sufficiently small. Let $F_n^\mu(x) = \Pr\{\hat{\mu}_{*,n} \leq x\}$ and $F_n^\tau(y) = \Pr\{\hat{\tau}_{*,n} \geq y\}$. Then, the following equations hold.*

$$\begin{aligned} \sum_{n=1}^{\infty} \int_{x=0}^{\mu_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(x, \mu^*(\epsilon))) dF_n^\mu(x) &= O(\epsilon^{-4}) \\ \sum_{n=1}^{\infty} \int_{y=1}^{\tau_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) dF_n^\tau(y) &= O(\epsilon^{-4}), \\ \sum_{n=1}^{\infty} \left(\int_{x=0}^{\mu_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(x, \mu^*(\epsilon))) dF_n^\mu(x) \right. \\ &\quad \cdot \left. \int_{y=1}^{\tau_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) dF_n^\tau(y) \right) = O(\epsilon^{-6}). \end{aligned}$$

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

$$\begin{aligned} &\int_{y=1}^{\tau_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau^*(\epsilon))) dF_n^\tau(y) \\ &= [\exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) F_n^\tau(y)]_{y=1}^{\tau_*(\epsilon)} + \int_{y=1}^{\tau_*(\epsilon)} -\frac{d \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon)))}{dy} F_n^\tau(y) dy. \end{aligned}$$

The first term can be bounded as

$$\begin{aligned}
 [\exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) F_n^\tau(y)]_{y=1}^{\tau_*^*(\epsilon)} &= F_n^\tau(\tau_*(\epsilon)) - \exp(n \mathcal{K}_{\mathcal{B}}(1, \tau_*(\epsilon))) F_n^\tau(1) \\
 &\leq F_n^\tau(\tau_*(\epsilon)) \\
 &= \Pr\{\hat{\tau}_{*,n} \geq \tau_* + \epsilon\} \\
 &\leq \exp(-n \mathcal{K}_{\mathcal{B}}(\tau_*(\epsilon), \tau_*))
 \end{aligned}$$

by using Cramer-Chernoff inequality (Fact 2). The second term is bounded as follows:

$$\begin{aligned}
 &\int_{y=1}^{\tau_*^*(\epsilon)} -\frac{d \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*^*(\epsilon)))}{dy} F_n^\tau(y) dy \\
 &= \int_{y=\tau_*(\epsilon)}^1 n \left(\ln \left(\frac{y}{\tau_*(\epsilon)} \right) - \ln \left(\frac{1-y}{1-\tau_*(\epsilon)} \right) \right) \cdot \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) F_n^\tau(y) dy
 \end{aligned}$$

Here we have

$$\begin{aligned}
 \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) F_n^\tau(y) &\leq \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*^*(\epsilon)) - n \mathcal{K}_{\mathcal{B}}(y, \tau_*^*)) \\
 &\leq \exp(-n f(\tau_*(\epsilon), \tau_*))
 \end{aligned}$$

by Proposition 1. Then,

$$\begin{aligned}
 &\int_{y=\tau_*(\epsilon)}^1 n \left(\ln \left(\frac{y}{\tau_*(\epsilon)} \right) - \ln \left(\frac{1-y}{1-\tau_*(\epsilon)} \right) \right) \cdot \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) F_n^\tau(y) dy \\
 &\leq \int_{y=\tau_*(\epsilon)}^1 n \left(\ln \left(\frac{y}{\tau_*(\epsilon)} \right) - \ln \left(\frac{1-y}{1-\tau_*(\epsilon)} \right) \right) \cdot \exp(-n f(\tau_*(\epsilon), \tau_*)) dy \\
 &= n \ln \left(\frac{1}{\tau_*(\epsilon)} \right) \exp(-n f(\tau_*(\epsilon), \tau_*^*)).
 \end{aligned}$$

Thus, we obtain

$$\begin{aligned}
 &\sum_{n=1}^{\infty} \int_{y=1}^{\tau_*^*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) dF_n^\tau(y) \\
 &\leq \sum_{n=1}^{\infty} \left(\exp(-n \mathcal{K}_{\mathcal{B}}(\tau_*(\epsilon), \tau_*)) + \ln \left(\frac{1}{\tau_*(\epsilon)} \right) n \exp(-n f(\tau_*(\epsilon), \tau_*)) \right) \\
 &\leq \underbrace{\frac{1}{\mathcal{K}(\tau_*(\epsilon), \tau_*)}}_{O(\epsilon^{-2}) \text{ (by Pinsker's inequality)}} + \underbrace{\ln \left(\frac{1}{\tau_*(\epsilon)} \right) \frac{\exp(f(\tau_*(\epsilon), \tau_*))}{f^2(\tau_*(\epsilon), \tau_*)}}_{O(\epsilon^{-4}) \text{ (by the definition of } f)} \\
 &= O(\epsilon^{-4}).
 \end{aligned}$$

Using the same argument yields

$$\sum_{n=1}^{\infty} \int_{x=0}^{\mu_*(\epsilon)} \exp(n \mathcal{K}(x, \mu_*(\epsilon))) dF_n^\mu(x) = O(\epsilon^{-4}).$$

Similarly, we have

$$\begin{aligned} & \sum_{n=1}^{\infty} \int_{x=0}^{\mu_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(x, \mu_*(\epsilon))) dF_n^\mu(x) \cdot \int_{y=1}^{\tau_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) dF_n^\tau(y) \\ & \leq \sum_{n=1}^{\infty} \left(\exp(-n \mathcal{K}(\mu_*(\epsilon), \mu_*)) + \ln \left(\frac{1}{1 - \mu_*(\epsilon)} \right) n \exp(-n f(\mu_*, \mu_*(\epsilon))) \right) \\ & \quad \cdot \left(\exp(-n \mathcal{K}_{\mathcal{B}}(\tau_*(\epsilon), \tau_*)) + \ln \left(\frac{1}{\tau_*(\epsilon)} \right) n \exp(-n f(\tau_*(\epsilon), \tau_*)) \right) \\ & \leq \frac{1}{\mathcal{K}_{\mathcal{B}}(\mu_*(\epsilon), \mu_*) + \mathcal{K}_{\mathcal{B}}(\tau_*(\epsilon), \tau_*)} \\ & \quad + \ln \left(\frac{1}{\tau_*(\epsilon)} \right) \frac{\exp(\mathcal{K}_{\mathcal{B}}(\mu_*(\epsilon), \mu_*) + f(\tau_*(\epsilon), \tau_*))}{(\mathcal{K}_{\mathcal{B}}(\mu_*(\epsilon), \mu_*) + f(\tau_*(\epsilon), \tau_*))^2} \\ & \quad + \ln \left(\frac{1}{1 - \mu_*(\epsilon)} \right) \cdot \frac{\exp(\mathcal{K}_{\mathcal{B}}(\tau_*(\epsilon), \tau_*) + f(\mu_*, \mu_*(\epsilon)))}{(\mathcal{K}_{\mathcal{B}}(\tau_*(\epsilon), \tau_*) + f(\mu_*, \mu_*(\epsilon)))^2} \\ & \quad + 2 \ln \left(\frac{1}{1 - \mu_*(\epsilon)} \right) \ln \left(\frac{1}{\tau_*(\epsilon)} \right) \cdot \frac{\exp(f(\mu_*, \mu_*(\epsilon)) + f(\tau_*(\epsilon), \tau_*))}{(f(\mu_*, \mu_*(\epsilon)) + f(\tau_*(\epsilon), \tau_*))^3} \\ & = O(\epsilon^{-6}). \end{aligned}$$

□

We are ready to give a main statement.

Lemma 3. *Let $\epsilon > 0$ be sufficiently small. Then, the following equality holds:*

$$\sum_{t=K+1}^T \Pr \left\{ u_*(t) < \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} = O(\epsilon^{-6}). \quad (3.15)$$

Proof. We have,

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

The first summation is bounded by

$$\begin{aligned}
 & \sum_{t=1}^T \Pr \left\{ u_*(t) < \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)}, \hat{\mu}_{*,N_*(t)} > \mu_*(\epsilon) \right\} \\
 &\leq \sum_{t=1}^T \Pr \left\{ \mathcal{K}_{\mathcal{B}}(\hat{\tau}_{*,N_*(t)}, \tau_*(\epsilon)) > \frac{\ln(t)}{N_*(t)}, \hat{\tau}_{*,N_*(t)} \geq \tau_*(\epsilon) \right\} \\
 &\leq \sum_{t=1}^T \sum_{n=1}^t \Pr \{ t < \exp(n \mathcal{K}_{\mathcal{B}}(\hat{\tau}_{*,n}, \tau_*(\epsilon))), \hat{\tau}_{*,n} \geq \tau_*(\epsilon), N_*(t) = n \} \\
 &= \sum_{n=1}^T \mathbb{E} \left[\sum_{t=n}^T \mathbb{I} \{ t < \exp(n \mathcal{K}_{\mathcal{B}}(\hat{\tau}_{*,n}, \tau_*(\epsilon))), N_*(t) = n \} \cdot \mathbb{I} \{ \hat{\tau}_{*,n} \geq \tau_*(\epsilon) \} \right] \\
 &\leq \sum_{n=1}^T \mathbb{E} [\exp(n \mathcal{K}_{\mathcal{B}}(\hat{\tau}_{*,n}, \tau_*(\epsilon))) \cdot \mathbb{I} \{ \hat{\tau}_{*,n} \geq \tau_*(\epsilon) \}] \\
 &= \sum_{n=1}^T \int_{y=1}^{\tau_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) dF_n^\tau(y) \\
 &= O(\epsilon^{-4}),
 \end{aligned}$$

where the last equality is by Lemma 2. In the same manner, we upper-bound the second summation as

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

and the third summation as

$$\begin{aligned}
 & \sum_{t=1}^T \Pr \left\{ u_*(t) < \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)}, \hat{\mu}_{*,N_*(t)} \leq \mu_*(\epsilon), \hat{\tau}_{*,N_*(t)} \geq \tau_*(\epsilon) \right\} \\
 & \leq \sum_{t=1}^T \Pr \left\{ \begin{aligned} & \mathcal{K}_{\mathcal{B}}(\hat{\mu}_{*,N_*(t)}, \mu_*(\epsilon)) + \mathcal{K}_{\mathcal{B}}(\hat{\tau}_{*,N_*(t)}, \tau_*(\epsilon)) > \frac{\ln(t)}{N_*(t)}, \\ & \hat{\mu}_{*,N_*(t)} \leq \mu_*(\epsilon), \hat{\tau}_{*,N_*(t)} \geq \tau_*(\epsilon) \end{aligned} \right\} \\
 & \leq \sum_{t=1}^T \sum_{n=1}^t \Pr \left\{ \begin{aligned} & t < \exp(n \mathcal{K}_{\mathcal{B}}(\hat{\mu}_{*,n}, \mu_*(\epsilon)) + n \mathcal{K}_{\mathcal{B}}(\hat{\tau}_{*,n}, \tau_*(\epsilon))), \\ & N_*(t) = n, \hat{\mu}_{*,n} \leq \mu_*(\epsilon), \hat{\tau}_{*,n} \geq \tau_*(\epsilon) \end{aligned} \right\} \\
 & = \sum_{n=1}^T \mathbb{E} \left[\sum_{t=n}^T \mathbb{I}\{t < \exp(n \mathcal{K}_{\mathcal{B}}(\hat{\mu}_{*,n}, \mu_*(\epsilon)) + n \mathcal{K}_{\mathcal{B}}(\hat{\tau}_{*,n}, \tau_*(\epsilon))), N_*(t) = n\} \right. \\
 & \quad \left. \cdot \mathbb{I}\{\hat{\mu}_{*,n} \leq \mu_*(\epsilon), \hat{\tau}_{*,n} \geq \tau_*(\epsilon)\} \right] \\
 & \leq \sum_{n=1}^T \mathbb{E} [\exp(n \mathcal{K}_{\mathcal{B}}(\hat{\mu}_{*,n}, \mu_*(\epsilon))) \cdot \mathbb{I}\{\hat{\mu}_{*,n} \leq \mu_*(\epsilon)\} \\
 & \quad \cdot \exp(n \mathcal{K}_{\mathcal{B}}(\hat{\tau}_{*,n}, \tau_*(\epsilon))) \cdot \mathbb{I}\{\hat{\tau}_{*,n} \geq \tau_*(\epsilon)\}] \\
 & = \sum_{n=1}^T \int_{x=0}^{\mu_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(x, \mu_*(\epsilon))) dF_n^\mu(x) \int_{y=1}^{\tau_*(\epsilon)} \exp(n \mathcal{K}_{\mathcal{B}}(y, \tau_*(\epsilon))) dF_n^\tau(y) \\
 & = O(\epsilon^{-6}),
 \end{aligned}$$

where the last inequality is by Lemma 2.

In summary, we obtain

$$\sum_{t=K+1}^T \Pr \left\{ u_*(t) < \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} = O(\epsilon^{-6}).$$

□

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

Lemma 4. *Let $\epsilon > 0$ be sufficiently small and*

$$\bar{N}_i(\epsilon) = \min \left\{ n : U \left(\mu_i + \epsilon, \tau_i - \epsilon, \frac{\ln(T)}{n}; \mathcal{K}_{\mathcal{B}} \right) < \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\}$$

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

$$\bar{N}_i(\epsilon) = \frac{1 + O(\epsilon)}{D_i} \ln(T) + 1. \tag{3.16}$$

Proof. Since

$$U\left(\mu_i + \epsilon, \tau_i - \epsilon, \frac{\ln(T)}{n}; \mathcal{K}_{\mathcal{B}}\right) < \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)}$$

is equivalent to

$$(a\mu_*(\epsilon), a\tau_*(\epsilon)) \notin \left\{(x, y) : \mathcal{K}_{\mathcal{B}}(\mu_i + \epsilon, x) + \mathcal{K}_{\mathcal{B}}(\tau_i - \epsilon, y) \leq \frac{\ln(t)}{n}\right\}$$

for all $a > 0$, $\bar{N}_i(\epsilon)$ is expressed as

$$\bar{N}_i(\epsilon) = \min \left\{ n : \min_{a>0} \left(\mathcal{K}_{\mathcal{B}}(\mu_i + \epsilon, a\mu_*(\epsilon)) + \mathcal{K}_{\mathcal{B}}(\tau_i - \epsilon, a\tau_*(\epsilon)) \right) > \frac{\ln(T)}{n} \right\}.$$

Because a that minimizes $\mathcal{K}_{\mathcal{B}}(\mu_i + \epsilon, a\mu_*(\epsilon)) + \mathcal{K}_{\mathcal{B}}(\tau_i - \epsilon, a\tau_*(\epsilon))$ lies² in an interval

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

there exists a constant h_i that satisfies³

$$\mathcal{K}_{\mathcal{B}}(\mu_i + \epsilon, a(\mu_* - \epsilon)) \geq \mathcal{K}_{\mathcal{B}}(\mu_i, a\mu_*) - h_i\epsilon$$

and

$$\mathcal{K}_{\mathcal{B}}(\tau_i - \epsilon, a(\tau_* + \epsilon)) \geq \mathcal{K}_{\mathcal{B}}(\tau_i, a\tau_*) - h_i\epsilon$$

for any appropriate a and sufficiently small ϵ . Thus,

$$\begin{aligned} \bar{N}_i(\epsilon) &\leq \min \left\{ n : \min_{a>0} (\mathcal{K}_{\mathcal{B}}(\mu_i, a\mu_*) + \mathcal{K}_{\mathcal{B}}(\tau_i, a\tau_*)) - 2h_i\epsilon > \frac{\ln(T)}{n} \right\} \\ &\leq \frac{\ln(T)}{D_i - 2h_i\epsilon} + 1 \\ &= \frac{\ln(T)}{D_i} \frac{1}{1 - 2h_i\epsilon D_i^{-1}} + 1 \\ &\leq \frac{1 + 4h_i\epsilon D_i^{-1}}{D_i} \ln(T) + 1 \end{aligned}$$

²Let $f_{\alpha,\beta}(a) = \mathcal{K}_{\mathcal{B}}(\alpha, \beta a)$. 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 Γ_i . 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 optimality of (μ_*, τ_*) .

³Let $g_1(\epsilon, a) = \mathcal{K}_{\mathcal{B}}(\mu_i + \epsilon, a\mu_*(\epsilon))$, $g_2(\epsilon, a) = \mathcal{K}_{\mathcal{B}}(\tau_i - \epsilon, a\tau_*(\epsilon))$. 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 $\epsilon \leq \epsilon_0$.

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

Lemma 5. *Let $\epsilon > 0$ be sufficiently small. Then, the following equality holds for any suboptimal action $i \neq *$:*

$$\sum_{t=K+1}^T \Pr \left\{ I(t) = i, u_*(t) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} = \frac{1 + O(\epsilon)}{D_i} \ln(T) + O(\epsilon^{-2}).$$

Proof. Since $I(t) = i \neq i^*$ implies $u_*(t) \leq u_i(t)$, we have

$$\begin{aligned} & \sum_{t=K+1}^T \mathbb{I} \left\{ I(t) = i, u_*(t) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} \\ & \leq \sum_{t=K+1}^T \mathbb{I} \left\{ I(t) = i, u_i(t) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} \\ & = \sum_{t=K+1}^T \sum_{n=1}^t \mathbb{I} \left\{ N_i(t) = n, I(t) = i, U \left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(t)}{n}; \mathcal{K}_{\mathcal{B}} \right) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} \\ & \leq \sum_{n=1}^T \sum_{t=n}^T \mathbb{I} \left\{ N_i(t) = n, I(t) = i, U \left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}; \mathcal{K}_{\mathcal{B}} \right) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} \\ & \leq \sum_{n=1}^T \mathbb{I} \left\{ U \left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}; \mathcal{K}_{\mathcal{B}} \right) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} \end{aligned} \quad (3.17)$$

Moreover, (3.17) is transformed as:

$$\begin{aligned} & \sum_{n=1}^T \mathbb{I} \left\{ U \left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}; \mathcal{K}_{\mathcal{B}} \right) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} \\ & \leq \bar{N}_i(\epsilon) + \sum_{n=\bar{N}_i(\epsilon)+1}^T \mathbb{I} \left\{ U \left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}; \mathcal{K}_{\mathcal{B}} \right) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} \\ & \leq \bar{N}_i(\epsilon) + \sum_{n=\bar{N}_i(\epsilon)+1}^{\infty} (\mathbb{I}\{\hat{\mu}_{i,n} \geq \mu_i + \epsilon\} + \mathbb{I}\{\hat{\tau}_{i,n} \leq \tau_i - \epsilon\}) \end{aligned} \quad (3.18)$$

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

$$U \left(\hat{\mu}_{i,n}, \hat{\tau}_{i,n}, \frac{\ln(T)}{n}; \mathcal{K}_{\mathcal{B}} \right) \leq U \left(\mu_i + \epsilon, \tau_i - \epsilon, \frac{\ln(T)}{n}; \mathcal{K}_{\mathcal{B}} \right) < \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)}.$$

Taking the expectation of (3.18) yields

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

where we used Hoeffding's inequality (Fact 1). Using Lemma 4 completes the proof. $\square$

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

Theorem 6. *Let π be the KL-UCB-SC. Then for any suboptimal action $i \neq *$, $T \geq 1$, and 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:

$$(3.19) \quad N_i(T) \leq 1 + \sum_{t=K+1}^T \mathbb{I} \left\{ u_*(t) < \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\} + \sum_{t=K+1}^T \mathbb{I} \left\{ I(t) = i, u_*(t) \geq \frac{\mu_*(\epsilon)}{\tau_*(\epsilon)} \right\}.$$

Then, the proof is immediately completed by taking the expectation and using Lemma 3 (for bounding the second term) and Lemma 5 (for bounding the third term). $\square$

Using Lemma 1, we immediately get following corollary, a upper bound of a regret of KL-UCB-SC as a function of budget B .

Corollary 6. *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 6 with $\epsilon = (\ln(B))^{-1/7}$ for sufficiently large B and Lemma 1.

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

□

3.4.3 Upper Bound of the Expected Regret for UCB-SC

In the exactly same manner, we can prove an upper bound of expected regret for UCB-SC. For UCB-SC, an action dissimilarity from the best action is defined as

$$D'_i = \min_{a > 0} (\mathcal{K}_{\mathcal{N}}(\mu_i, a\mu_*) + \mathcal{K}_{\mathcal{N}}(\tau_i, a\tau_*)).$$

Unlike to the that for KL-UCB-SC, it can be expressed in a closed-form.

Lemma 6.

$$D'_i = \frac{2\tau_*^2}{\mu_*^2 + \tau_*^2} \tau_i^2 \Delta_i^2.$$

Proof. From the definition of D'_i ,

$$\begin{aligned}
 D_i &= \min_a (\mathcal{K}_{\mathcal{B}}(\mu_i, a\mu_*) + \mathcal{K}_{\mathcal{B}}(\tau_i, a\tau_*)) \\
 &\geq \min_a (2(a\mu_* - \mu_i)^2 + 2(a\tau_* - \tau_i)^2) \\
 &= \frac{2(\mu_i\tau_* - \mu_*\tau_i)^2}{\mu_*^2 + \tau_*^2} \\
 &= \frac{2(\Delta_i\tau_i\tau_*)^2}{\mu_*^2 + \tau_*^2},
 \end{aligned}$$

where the 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. □

Note that by Pinsker's inequality (Fact 3), $D_i > D'_i$ holds for any $i \neq *$. The upper bound of expected regret for UCB-SC is proved as follows.

Theorem 7. *Let π be the UCB-SC. Then the regret is upper-bounded as:*

$$\begin{aligned}\tilde{R}_\pi(B) &= \sum_{i: \Delta_i > 0} \frac{\tau_i \Delta_i \ln(B)}{D'_i} + o(\ln(B)) \\ &= \sum_{i: \Delta_i > 0} \frac{1}{2} \left(\left(\frac{\mu_*}{\tau_*} \right)^2 + 1 \right) \frac{\ln(B)}{\tau_i \Delta_i} + o(\ln(B))\end{aligned}$$

Note that the upper bound proved in Theorem 7 holds even if the policy was KL-UCB-SC, since $D_i > D'_i$.

3.5 Lower Bound of Regret

Following a similar discussion to the one in [12], 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 2 (Strong consistency). *A policy is strongly consistent if, for any suboptimal action i and $p > 0$*

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E}[N_i(T)]}{T^p} = 0$$

holds for any distributions.

We will write a collection of strong consistent policy as Π^* .

Theorem 8. *Let the rewards and costs be drawn from Bernoulli distributions with parameters $\{\mu_i, \tau_i\}_{i \in A}$. 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. For any suboptimal action i and small positive constant $\epsilon > 0$, there exist distribution parameters μ'_i and τ'_i such that $\mu'_i \geq \mu_i$, $\tau'_i < \tau_i$,

$$\frac{\mu'_i}{\tau'_i} > \frac{\mu_*}{\tau_*} > \frac{\mu_i}{\tau_i},$$

and

$$\begin{aligned}\mathcal{K}_B(\mu_i, \mu'_i) + \mathcal{K}_B(\tau_i, \tau'_i) &\leq (1 + \epsilon) \min_{a > 0} (\mathcal{K}_B(\mu_i, a\mu_*) + \mathcal{K}_B(\tau_i, a\tau_*)) \\ &= (1 + \epsilon) D_i.\end{aligned}\tag{3.20}$$

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 $*$ but action i .

Let $p_X(x)$ and $p_Y(x)$ 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 $q_Y(x)$ be the PDFs of $\mathcal{B}(\mu'_i)$ and $\mathcal{B}(\tau'_i)$ respectively.

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

$$\begin{aligned} \text{ELLR}(n) &= \ln \left(\frac{p_X(X_{i,1}) \cdots p_X(X_{i,n}) \cdot p_Y(Y_{i,1}) \cdots p_Y(Y_{i,n})}{q_X(X_{i,1}) \cdots q_X(X_{i,n}) \cdot q_Y(Y_{i,1}) \cdots q_Y(Y_{i,n})} \right) \\ &= \sum_{s=1}^n \left(\ln \frac{p_X(X_{i,s})}{q_X(X_{i,s})} + \ln \frac{p_Y(Y_{i,s})}{q_Y(Y_{i,s})} \right) \end{aligned}$$

Note that

$$\mathbb{E} \left[\ln \frac{p_X(X_{i,s})}{q_X(X_{i,s})} + \ln \frac{p_Y(Y_{i,s})}{q_Y(Y_{i,s})} \right] = \mathcal{K}_{\mathcal{B}}(\mu_i, \mu'_i) + \mathcal{K}_{\mathcal{B}}(\tau_i, \tau'_i)$$

holds.

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

$$\Pr'\{E_T\} = \mathbb{E} [\mathbb{I}\{E_T\} \exp(-\text{ELLR}(N_i(T)))] \quad (3.21)$$

holds.

Define two events

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

and

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

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

$$S(T) = \frac{(1 - \epsilon) \ln(T)}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu'_i) + \mathcal{K}_{\mathcal{B}}(\tau_i, \tau'_i)}.$$

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. (3.21) and the definition of A_T and B_T , we have

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

Then,

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

By using the Markov's inequality (Fact 4), we obtain

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

On the modified bandit problem, the optimal action is not $*$ but i . Therefore, the strong consistency of the policy requires

$$\mathbb{E}'[T - N_i(T)] = \mathbb{E}'\left[\sum_{j \neq i} N_j(T)\right] = o(T^p)$$

for any $p > 0$. By using this we obtain

$$\Pr\{A_T, B_T\} = o(1). \quad (3.22)$$

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

$$\begin{aligned}\Pr\{A_T, \bar{B}_T\} &= \Pr\{N_i(T) \leq S(T), \text{ELLR}(N_i(T)) > (1 - \epsilon') \ln(T)\} \\ &\leq \Pr\left\{\max_{s \leq S(T)} \text{ELLR}(s) > (1 - \epsilon') \ln(T)\right\} \\ &= \Pr\left\{\frac{1}{S(T)} \max_{s \leq S(T)} \text{ELLR}(s) > \frac{1 - \epsilon'}{1 - \epsilon} (\mathcal{K}_{\mathcal{B}}(\mu_i, \mu'_i) + \mathcal{K}_{\mathcal{B}}(\tau_i, \tau'_i))\right\}\end{aligned}$$

Using Fact 5, and the expectation of $\text{ELLR}(s)$,

$$\lim_{T \rightarrow \infty} \frac{1}{S(T)} \max_{s \leq S(T)} \text{ELLR}(s) \rightarrow \mathcal{K}_{\mathcal{B}}(\mu_i, \mu'_i) + \mathcal{K}_{\mathcal{B}}(\tau_i, \tau'_i) \quad \text{a.s.}$$

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

$$\Pr\{A_T, \bar{B}_T\} = o(1). \quad (3.23)$$

By (3.22) and (3.23), we have

$$\Pr\{A_T\} = \Pr\{A_T, B_T\} + \Pr\{A_T, \bar{B}_T\} = o(1). \quad (3.24)$$

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

$$\begin{aligned}
 \mathbb{E}[N_i(T)] &\geq S(T) \cdot \Pr\{N_i(T) \geq S(T)\} \\
 &= S(T)(1 - \Pr\{A_T\}) \\
 &= \frac{(1 - \epsilon) \ln(T)}{\mathcal{K}_{\mathcal{B}}(\mu_i, \mu'_i) + \mathcal{K}_{\mathcal{B}}(\tau_i, \tau'_i)} (1 - o(1)) \\
 &\geq \frac{1 - \epsilon}{1 + \epsilon} (1 - o(1)) \frac{\ln(T)}{D_i} \\
 &\geq (1 - o(1)) \frac{\ln(T)}{D_i},
 \end{aligned}$$

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

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 .

Corollary 7. *Let the rewards and costs be drawn from Bernoulli distributions with parameters $\{\mu_i, \tau_i\}_{i \in A}$. Then the regret is lower-bounded as:*

$$\inf \left\{ \tilde{R}_{\pi}(B) \mid \pi \in \Pi^* \right\} = \sum_{i: \Delta_i > 0} \frac{\tau_i \Delta_i \ln(B)}{D_i} - o(\ln(B)).$$

Proof. Similar to the proof of Theorem 6, using Theorem 8 and Lemma 1, we find that

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

$\square$

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

Corollary 8. *On a 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. More formally, Let π be the KL-UCB-SC, Then the following equation is true.

$$\lim_{B \rightarrow \infty} \frac{R_\pi(B)}{\inf\{R_{\pi'}(B) \mid \pi' \in \Pi^*\}} = 1.$$

3.5.1 Comparison with other policies

For a budgeted multi-armed bandits with stochastic costs, a number of policies, such as UCB-BV1 [13] and BTS [17] 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 [17]

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

UCB-BV1 [13] ($\lambda = \tau_{\min}$)

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

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. (3.2).

These bounds and the bound we derived for KL-UCB-SC are not naturally comparable. But by comparing bound for UCB-SC (Theorem 7) with those in Remark 1, the regret bound improves at least by a factor of four on the best known one (proved for BTS [17]). Moreover, as described earlier, in comparison of the proved upper bound for KL-UCB-SC and UCB-SC, the upper bound for KL-UCB-SC is always smaller. These facts immediately conclude that the upper bound proved for KL-UCB-SC was essentially better than shown in Remark 1. 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.

3.6 Experiments

3.6.1 Setup

We conduct a numerical simulation using simple models in order to confirm the characteristics of UCB-SC compared to KL-UCB-SC and the effectiveness of UCB-SC and its variant compared to other existing policies. We compared the following five policies in our experiments: KL-UCB-SC, KL-UCB-SC+, BTS [17], PD-BwK [28] and UCB-BV1 [13]. KL-UCB-SC+ is a version of KL-UCB-SC that uses $\ln(t/N_i(t-1))$ instead of $\ln(t)$, which is inspired by the KL-UCB+ policy for the classical multi-armed bandits [18]. The introduction of the denominator $N_i(t-1)$ does not change the asymptotic property because for any suboptimal action i , $N_i(t-1) = O(\ln t)$ holds almost surely and thus $\ln(t/N_i(t-1))$ is still $O(\ln t)$, while it generally improves a finite-time performance [7, 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. But for BTS, the last two 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

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

UCB-BV1

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

where $[x]_{\geq \theta} = \max\{x, \theta\}$, $[x]_{\leq \theta} = \min\{x, \theta\}$. The parameters of PD-BwK and UCB-BV1 are set to $C_{\text{rad}} = 0.25 \ln(KB)$ and $\lambda = \tau_{\min}$ respectively⁴.

Simulations are done using two scenarios (reward/cost distributions) shown in Table 3.1. The reward and cost means are set to around 0.5 in scenarios M and 0.75 in scenarios H. Note that action i in the scenarios M and H has

⁴To guarantee the regret upper bound shown in [28], setting $C_{\text{rad}} = \Theta(\ln(KB))$ is necessary for PD-BwK and $C_{\text{rad}} = 0.25 \ln(KB)$ is the setting used in the comparison experiment in [17]. Setting $\lambda = \tau_{\min}$ is optimal for UCB-BV1 in terms of the regret upper bound shown in [13].

Table 3.1: Pairs $(X_i(t), Y_i(t))$ of distributions for actions i

(a) Scenario M		(b) Scenario H	
i	$(X_i(t), Y_i(t))$	i	$(X_i(t), Y_i(t))$
1	$(\mathcal{B}(0.5), \mathcal{B}(0.5))$	1	$(\mathcal{B}(0.75), \mathcal{B}(0.75))$
2	$(\mathcal{B}(0.45), \mathcal{B}(0.45))$	2	$(\mathcal{B}(0.675), \mathcal{B}(0.675))$
3	$(\mathcal{B}(0.55), \mathcal{B}(0.55))$	3	$(\mathcal{B}(0.825), \mathcal{B}(0.825))$
4	$(\mathcal{B}(0.45), \mathcal{B}(0.55))$	4	$(\mathcal{B}(0.675), \mathcal{B}(0.825))$
5	$(\mathcal{B}(0.55), \mathcal{B}(0.45))$	5	$(\mathcal{B}(0.825), \mathcal{B}(0.675))$
6	$(\mathcal{B}(0.4), \mathcal{B}(0.6))$	6	$(\mathcal{B}(0.6), \mathcal{B}(0.9))$
7	$(\mathcal{B}(0.4), \mathcal{B}(0.5))$	7	$(\mathcal{B}(0.6), \mathcal{B}(0.75))$
8	$(\mathcal{B}(0.4), \mathcal{B}(0.4))$	8	$(\mathcal{B}(0.6), \mathcal{B}(0.6))$
9	$(\mathcal{B}(0.5), \mathcal{B}(0.6))$	9	$(\mathcal{B}(0.75), \mathcal{B}(0.9))$
10	$(\mathcal{B}(0.5), \mathcal{B}(0.4))$	10	$(\mathcal{B}(0.75), \mathcal{B}(0.6))$
11	$(\mathcal{B}(0.6), \mathcal{B}(0.6))$	11	$(\mathcal{B}(0.9), \mathcal{B}(0.9))$
12	$(\mathcal{B}(0.6), \mathcal{B}(0.5))$	12	$(\mathcal{B}(0.9), \mathcal{B}(0.75))$
$13 = i^*$	$(\mathcal{B}(0.6), \mathcal{B}(0.4))$	$13 = i^*$	$(\mathcal{B}(0.9), \mathcal{B}(0.6))$

the same reward-per-cost μ_i/τ_i for each i but both μ_i and τ_i of scenario H are 1.5 times larger than those of scenario M.

All experiments were run on a computer with Intel(R) Core(TM) i5 3.2 GHz CPU and 16 GB 1600 MHz DDR3 memory.

3.6.2 Results

Average Regret

The empirical regret averaged over 100 runs on scenario M and H with $B = 10\,000$ are shown in Fig. 3.1, where the empirical regret $\hat{R}(B)$ is calculated as

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

UCB-SC is outperformed by KL-UCB-SC but the difference is small in scenario M compared to that in scenario H, which matches the difference of theoretical bounds considering tightness of Pinsker's Inequality. UCB-SC+ performs better than KL-UCB-SC, and performs as well as KL-UCB-SC+ in scenario M but not so well as KL-UCB-SC+ in scenario H.

Computation Time

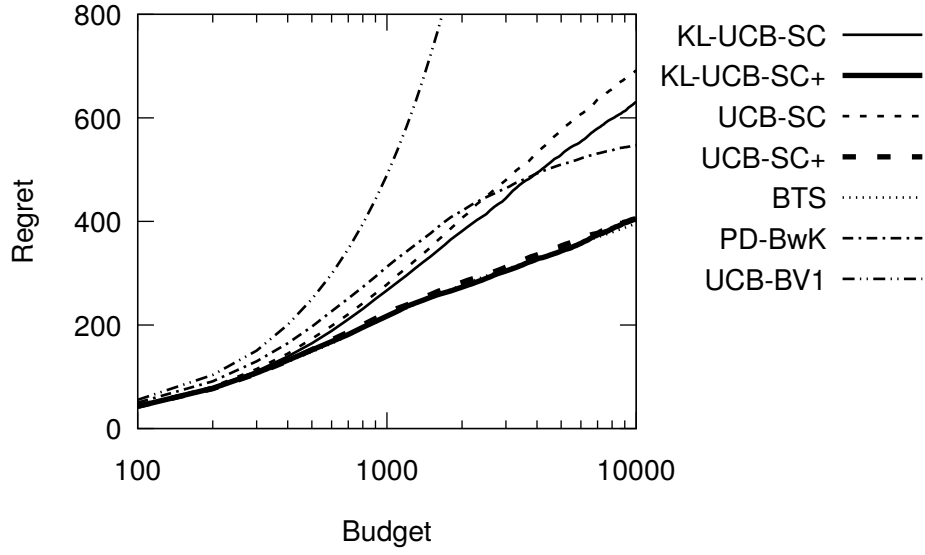
In Table 3.2, we show computation time per round of each policy averaged over 100 independent runs on scenarios M and H. Note that computation time per round linearly depends on the number of actions for all the policies. UCB-SC(+) runs more than 46 000 times faster than KL-UCB-SC(+) on both scenarios. Furthermore, UCB-SC and UCB-SC+, together with PD-BwK and UCB-BV1, can be categorized into the fastest policies. Among them, UCB-SC+ achieves the smallest average regret for both scenarios.

Table 3.2: Computation time [μ s] per round averaged over 100 runs.

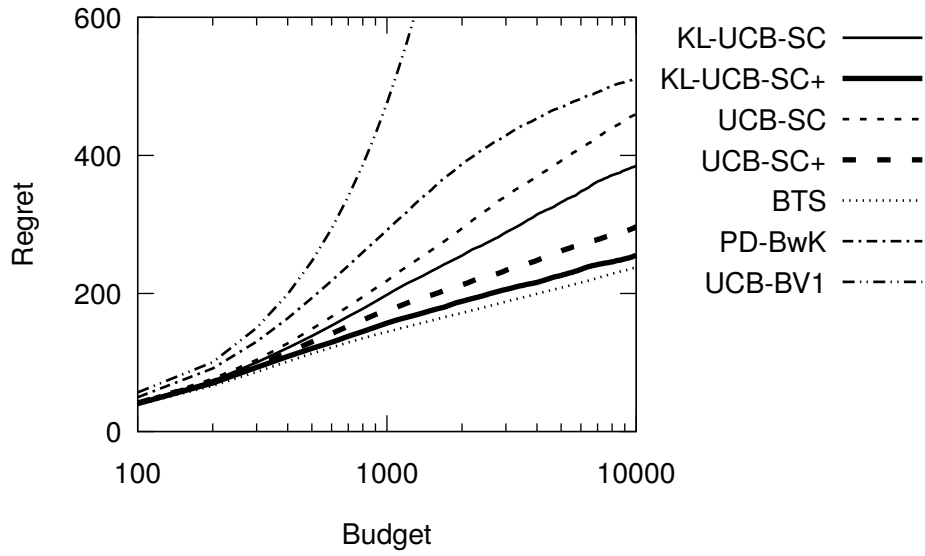
Policy	Scenario M	Scenario H
UCB-SC	1.029	1.021
UCB-SC+	1.030	1.020
KL-UCB-SC	50 196.465	46 202.071
KL-UCB-SC+	51 081.726	48 109.738
UCB-BV1	0.587	0.588
PD-BwK	0.750	0.743
BTS	4.893	4.883

3.7 Summary

In this chapter, we proposed a policy named KL-UCB-SC for budgeted multi-armed bandit problem with stochastic action costs. Moreover, we developed a policy UCB-SC from KL-UCB-SC by analogy with the relation between UCB and KL-UCB on the classical multi-armed bandit problem. Unlike KL-UCB-SC, UCB-SC has an index of closed form that can be calculated immediately. We derived a regret upper bound for KL-UCB-SC 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 bandit problem. Furthermore, we demonstrated that the performance of KL-UCB-SC+, which is a tuned variant of KL-UCB-SC, is comparable to that of the state-of-the-art policies in numerical experiments. As for UCB-SC+, a heuristic variant of UCB-SC, performs best in terms of regret among the four fastest policies, UCB-BV1, PD-BwK, UCB-SC and UCB-SC+, according to our numerical simulations.



(a) Scenario M



(b) Scenario H

Figure 3.1: The empirical regret curves averaged over 100 independent runs with $B = 10\,000$ for scenarios M and H.

Chapter 4

Matching-Selection Bandit Problem

4.1 Problem Definition

Matching-selection bandit problem can be described as a combinatorial semi-bandit problem whose action set is represented as an edges of complete bi-partite graph $K_{M,N}$. For simplicity, we assume that $N \geq M$. We will write an edge as a tuple of two integer, that is

$$A = \{(i, j) \mid i \in [M], j \in [N]\}.$$

Moreover, any collection of a subset of action that satisfy matching constraint as $\mathcal{M} \subset \mathfrak{P}(A)$, where $\mathfrak{P}$ is a power set operator.

At each round $t = 1, 2, \dots$, a player chooses a subset of action set $I(t) \in A$ so as to satisfy maximal matching constraint. Then, a player receives reward $X_a(t) \in [0, 1]$ for each $a \in I(t)$. Rewards $X_a(t)$ is generated from a correspond distribution ν_a identically. We denote a mean of reward from action a as μ_a .

If a decision $I(t)$ satisfies maximal matching constraint, $I(t)$ behaves like a M -dimensional vector of $[N]$. We introduce a subscript notation $I_i(t)$ of $I(t)$, which implies

$$I(t) = \{(i, I_i(t)) \mid i \in [M]\}.$$

These notations are used interchangeably according to the context.

Therefore, a history is defined as

$$\mathcal{H}(t) = \left\{ (I(s), X_{(1, I_1(s))}(s), \dots, X_{(M, I_M(s))}(s)) \right\}_{s=1}^t$$

and a policy is defined as an infinite sequence of map $\pi = \pi_1, \pi_2, \dots$, where

$$\pi_t: ([N]^M \times [0, 1]^M)^{t-1} \rightarrow \mathcal{M}.$$

Literally,

$$I(t) = \pi_t(\mathcal{H}(t-1)).$$

The pseudo-code of matching-selection bandit problem is in Algorithm 3.

Algorithm 3 Matching-Selection Bandits

Require: A : Action set.

```

1: for  $t = 1, \dots, T$  do
2:   Choose  $I(t) \in \mathcal{M} \subseteq \mathfrak{P}(A)$ .
3:   for all  $a \in I(t)$  do
4:     Receive a reward  $X_a(t)$ .
5:   end for
6: end for

```

The goal of player is maximizing a cumulative reward

$$G_\pi(T) = \sum_{t=1}^T \sum_{a \in I(t)} X_a(t).$$

In order to assess the performance of a policy, we further introduce expected regret, defined as an expected difference between the cumulative reward by the best policy π^* and that by the given policy π , literally

$$R_\pi(T) = \mathbb{E}[G_{\pi^*}(T) - G_\pi(T)].$$

The best policy π^* is defined as

$$\pi^* = \arg \max_{\pi \in \Pi} \mathbb{E}[G_\pi(T)]$$

where Π be the any collection of policies. Same as the classical one, the best policy is the policy who always chooses the best set I^* , defined as

$$I^* = \arg \max_{I \in \mathcal{M}} \mu_I,$$

where

$$\mu_I = \sum_{a \in I} \mu_a.$$

We can have an equivalent reduced form of regret

$$R_\pi(T) = \mathbb{E} \left[\sum_{t=1}^T \Delta_{I(t)} \right],$$

where $\Delta_I = \mu_{I^*} - \mu_I$ for any $I \in \mathcal{M}$. Moreover, we denote $\Delta_{\min} = \min\{\Delta_I \mid I \in \mathcal{M}, \Delta_I > 0\}$ and $\Delta_{\max} = \max\{\Delta_I \mid I \in \mathcal{M}\}$.

4.2 LLR (Learning with Linear Rewards)

In this section, we describe a policy LLR [20], whose upper bound of expected regret we improve. The original one was designed for more general problem, but here we describe the specific version for the matching-selection bandit problem. It is known that a little modification incorporated in to CUCB [21] enables a better upper bound of expected regret to be proved, thus we consider a version of the proof in [21]. Same as other policies for many kinds of multi-armed bandit problem, LLR chooses a matching who maximizes evaluation index $u_I(t)$, i.e.

$$I(t) = \arg \max_{I \in \mathcal{M}} u_I(t).$$

In order to describe a definition of $u_I(t)$, we introduce some notations. $N_a(t)$ is a number of decisions which contains an action a up to time t , defined as

Let

$$N_a(t) = \sum_{s=1}^t \mathbb{I}\{a \in I(s)\}.$$

be the number of decisions which contains an action $a \in A$ up to time t . We define an estimator of μ_a in two kind of forms.

$$\hat{\mu}_{a,n} = \frac{1}{n} \sum_{s: N_a(s) \leq n, a \in I(s)} X_a(s)$$

and

$$\hat{\mu}_a(t) = \mu_a(N_a(t)) = \frac{1}{N_a(t)} \sum_{s \leq t, a \in I(s)} X_a(s)$$

First one specifies a number of observation used for estimate, second one specifies the time t . These notations are used interchangeably according to the context.

An action set evaluation index $u_I(t)$ for LLR is defined as

$$u_I(t) = \sum_{a \in I} u_a(t),$$

where

$$u_a(t) = \hat{\mu}_a(t-1) + \sqrt{\frac{\eta \ln(t)}{N_a(t-1)}}.$$

Here we employed a parametrized version. The original LLR uses $\eta = 3$.

A set of action $I(t)$ for LLR can be found by solving a maximum weight bipartite matching problem. A maximum weight bipartite matching problem with non-negative real weight can be solved $O(N^3)$ time by Hungarian method [29].

On a general combinatorial semi-bandit problem, an upper bound of expected regret LLR were proved.

Theorem 9 ([21]). *On a general combinatorial semi-bandit problem, let K be the number of action and m be the maximum cardinality of the set of action that satisfies given constraint. Then for LLR,*

$$R_\pi(T) = O(m^2 K \ln(T))$$

holds where T is the number of decisions.

On a matching-selection bandit problem over complete bipartite graph $K_{M,N}$, clearly we find that $K = MN$ and $m = M$. Hence, we immediately have following corollary.

Corollary 9. *On a matching-selection bandit problem over $K_{M,N}$, let π be LLR. Then,*

$$R_\pi(T) = O(M^3 N \ln(T))$$

holds where T is the number of decisions.

4.3 Regret Analysis

This section proves a refined upper bound of regret when a player obeys LLR on a matching-selection bandit problem on complete bipartite graph $K_{M,N}$. In order to show the main statement, we prepare some propositions and lemmas.

Proposition 3. *Assume that $a_1 \geq a_2 \geq \dots \geq a_K$ and $0 < b_1 \leq b_2 \leq \dots \leq b_K$ for some $K \in \mathbb{N}$. Then*

$$\sum_{k=1}^K a_k x_k \leq a_1 b_1 + \sum_{k=2}^K a_k (b_k - b_{k-1})$$

holds under the constraints $x_k \geq 0$ and

$$\sum_{\ell=1}^k x_\ell \leq b_k$$

for $k = 1, \dots, K$.

Proof. This is trivial since

$$\begin{aligned}
& (a_1 b_1 + a_2(b_2 - b_1) + \cdots + a_K(b_K - b_{K-1})) - (a_1 x_1 + a_2 x_2 + \cdots + a_K x_K) \\
&= (a_1 - a_2)(b_1 - x_1) + (a_2 - a_3) \left(b_2 - \sum_{\ell=1}^2 x_\ell \right) \\
&\quad + \cdots + (a_{K-1} - a_K) \left(b_{K-1} - \sum_{\ell=1}^{K-1} x_\ell \right) + a_K \left(b_K - \sum_{\ell=1}^K x_\ell \right) \\
&\geq 0
\end{aligned}$$

holds. $\square$

Proposition 4. For any real sequence $a_1, a_2, \dots, a_n$ with $1 \geq a_1 \geq a_2 \geq \cdots \geq a_n > 0$ and $n \geq 2$,

$$\frac{1}{a_1} + \sum_{k=2}^n a_k \left(\frac{1}{a_k^2} - \frac{1}{a_{k-1}^2} \right) + 1 \leq \frac{2}{a_n} \quad (4.1)$$

holds.

Proof. This proof is done by the mathematical induction on n . In case of $n = 2$, Eq. (4.1) will be

$$\frac{2}{a_2} - \left(\frac{1}{a_1} + a_2 \left(\frac{1}{a_2^2} - \frac{1}{a_1^2} \right) + 1 \right) = a_2 \left(\frac{1}{a_2} - \frac{1}{a_1} \right) + a_1 - 1 \geq 0.$$

Let Eq. (4.1) holds for some $n = k$, then for $n = k + 1$, we can find that

$$\begin{aligned}
& \frac{1}{a_1} + a_2 \left(\frac{1}{a_2^2} - \frac{1}{a_1^2} \right) + \cdots + a_{k+1} \left(\frac{1}{a_{k+1}^2} - \frac{1}{a_k^2} \right) + 1 \\
&\leq \frac{2}{a_k} + a_{k+1} \left(\frac{1}{a_{k+1}^2} - \frac{1}{a_k^2} \right) \\
&= \frac{2}{a_{k+1}} \frac{a_{k+1}}{a_k} + \frac{1}{a_{k+1}} - \frac{1}{a_{k+1}} \left(\frac{a_{k+1}}{a_k} \right)^2 \\
&= \frac{2}{a_{k+1}} - \frac{1}{a_{k+1}} \left(\frac{a_{k+1}}{a_k} - 1 \right)^2 \leq \frac{2}{a_{k+1}}
\end{aligned}$$

holds. Therefore, Eq. (4.1) holds for $n = k + 1$. $\square$

Lemma 7. Let Z be a natural number. For any policy,

$$\sum_{t=1}^T \mathbb{I} \left\{ \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \sqrt{\frac{M}{Z}} \right\} \leq 8M^{4/3}NZ \quad (4.2)$$

holds.

Proof. If Eq. (4.2) is true, one of following must be true:

$$\exists a \in I(t), N_a(t-1) \leq 4M^{1/3}Z \quad (4.3)$$

$$\#\{a \mid N_a(t-1) < 4MZ\} \geq M^{2/3}. \quad (4.4)$$

This can be verified as

$$\begin{aligned} & \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} \\ &= \sum_{a \in I(t), N_a(t-1) < 4MZ} \frac{1}{\sqrt{N_a(t-1)}} + \sum_{a \in I(t), N_a(t-1) \geq 4MZ} \frac{1}{\sqrt{N_a(t-1)}} \\ &\leq M^{2/3} \frac{1}{\sqrt{4M^{1/3}Z}} + (M - M^{2/3}) \frac{1}{\sqrt{4MZ}} \\ &= \frac{1}{2} \sqrt{\frac{M}{Z}} + \frac{1}{2} \sqrt{\frac{M}{Z}} - \frac{M^{1/6}}{2\sqrt{Z}} \\ &\leq \sqrt{\frac{M}{Z}}. \end{aligned}$$

Eq. (4.3) occurs at most $4M^{1/3}Z \cdot MN = 4M^{4/3}NZ$ times and Eq. (4.4) occurs at most $4MZ \cdot MN/M^{2/3} = 4M^{4/3}NZ$ times. Therefore, one of them hold true at most $8M^{4/3}NZ$ times. $\square$

Lemma 8. Let $Z_I = \lceil 2\eta \ln(T)/\Delta_I^2 \rceil$. Then,

$$\sum_{t=1}^T \Delta_{I(t)} \mathbb{I} \left\{ \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \sqrt{\frac{M}{Z_{I(t)}}} \right\} \leq \frac{32\eta M^{7/3}N}{\Delta_{\min}} \ln(T).$$

Proof. Let $I^{(k)}$ denote the k th largest matching $I \in \mathcal{M}$ ordered by Δ_I . Let K be the number of matchings I in $\mathcal{M}$ with $\Delta_I > 0$, that is,

$$\Delta_{\max} = \Delta_{I^{(1)}} \geq \Delta_{I^{(2)}} \geq \dots \geq \Delta_{I^{(K)}} = \Delta_{\min} > 0$$

and

$$\Delta_{I^{(K+1)}} = \dots = \Delta_{I^{(|\mathcal{M}|)}} = 0.$$

Note that

$$Z_{I^{(1)}} \leq Z_{I^{(2)}} \leq \dots \leq Z_{I^{(K)}}$$

holds by the definition of Z_I .

Let $a_k = \Delta_{I^{(k)}}$, $b_k = 8M^{4/3}NZ_{I^{(k)}}$ and

$$x_k = \sum_{t=1}^T \mathbb{I} \left\{ I(t) = I^{(k)}, \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \sqrt{\frac{M}{Z_{I^{(k)}}}} \right\}.$$

Then by Lemma 7, we can verify that

$$\begin{aligned} \sum_{\ell=1}^k x_\ell &= \sum_{\ell=1}^k \sum_{t=1}^T \mathbb{I} \left\{ I(t) = I^{(k)}, \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \sqrt{\frac{M}{Z_{I^{(k)}}}} \right\} \\ &\leq \sum_{t=1}^T \mathbb{I} \left\{ \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \sqrt{\frac{M}{Z_{I^{(k)}}}} \right\} \\ &\leq 8M^{4/3}NZ_{I^{(k)}} = b_k. \end{aligned}$$

Thus, by Proposition 3, we have

$$\begin{aligned} &\sum_{t=1}^T \Delta_{I(t)} \mathbb{I} \left\{ \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \sqrt{\frac{M}{Z_{I(t)}}} \right\} \\ &= \sum_{k=1}^K \Delta_{I^{(k)}} \sum_{t=1}^T \mathbb{I} \left\{ I(t) = I^{(k)}, \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \sqrt{\frac{M}{Z_{I^{(k)}}}} \right\} \\ &\leq \Delta_{I^{(1)}} \cdot 8M^{4/3}NZ_{I^{(1)}} + \sum_{k=2}^K \Delta_{I^{(k)}} (8M^{4/3}NZ_{I^{(k)}} - 8M^{4/3}NZ_{I^{(k-1)}}). \end{aligned}$$

By the definition of Z_I ,

$$\begin{aligned}
& \Delta_{I^{(1)}} \cdot 8M^{4/3}NZ_{I^{(1)}} + \sum_{k=2}^K \Delta_{I^{(k)}}(8M^{4/3}NZ_{I^{(k)}} - 8M^{4/3}NZ_{I^{(k-1)}}) \\
&= 8M^{4/3}N \left(\Delta_{I^{(1)}} \left\lceil \frac{2\eta \ln(T)}{\Delta_{I^{(1)}}^2} \right\rceil \right. \\
&\quad \left. + \sum_{k=1}^K \left(\Delta_{I^{(k)}} \left\lceil \frac{2\eta \ln(T)}{\Delta_{I^{(k)}}^2} \right\rceil - \Delta_{I^{(k-1)}} \left\lceil \frac{2\eta \ln(T)}{\Delta_{I^{(k-1)}}^2} \right\rceil \right) \right) \\
&= 8M^{4/3}N \left(\sum_{k=1}^{K-1} (\Delta_{I^{(k)}} - \Delta_{I^{(k+1)}}) \left\lceil \frac{2\eta \ln(T)}{\Delta_{I^{(k)}}^2} \right\rceil + \Delta_{I^{(K)}} \left\lceil \frac{2\eta \ln(T)}{\Delta_{I^{(K)}}^2} \right\rceil \right) \\
&\leq 8M^{4/3}N \left(\sum_{k=1}^{K-1} (\Delta_{I^{(k)}} - \Delta_{I^{(k+1)}}) \frac{2\eta \ln(T)}{\Delta_{I^{(k)}}^2} + \Delta_{I^{(K)}} \frac{2\eta \ln(T)}{\Delta_{I^{(K)}}^2} + \Delta_{I^{(1)}} \right) \\
&\leq 16\eta M^{7/3}N \ln(T) \left(\frac{1}{\Delta_{I^{(1)}}} + \sum_{k=2}^K \Delta_{I^{(k)}} \left(\frac{1}{\Delta_{I^{(k)}}^2} - \frac{1}{\Delta_{I^{(k-1)}}^2} \right) + 1 \right).
\end{aligned}$$

Using Proposition 4, we have

$$\begin{aligned}
& 16\eta M^{7/3}N \ln(T) \left(\frac{1}{\Delta_{I^{(1)}}} + \sum_{k=2}^K \Delta_{I^{(k)}} \left(\frac{1}{\Delta_{I^{(k)}}^2} - \frac{1}{\Delta_{I^{(k-1)}}^2} \right) + 1 \right) \\
&\leq \frac{32\eta M^{7/3}N \ln(T)}{\Delta_{I^{(K)}}} = \frac{32\eta M^{7/3}N \ln(T)}{\Delta_{\min}}.
\end{aligned}$$

□

Up to here, we are ready to show the main statement.

Theorem 10. *Let π be LLR with $\eta > 2$,*

$$R_\pi(T) \leq \frac{32\eta M^{7/4}N \ln(T)}{\Delta_{\min}} + \frac{MN\Delta_{\max}}{\eta - 2}$$

holds.

Proof. The event that selection of $I(t) \neq I^*$ can be decomposed into following

three events.

$$\begin{aligned}
& \{I(t) \neq I^*\} \\
& \subseteq \left\{ \exists a \in I^*, \hat{\mu}_a(t-1) \leq \mu_a - \sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} \right\} \\
& \cup \left\{ \exists a \notin I^*, \hat{\mu}_a(t-1) \geq \mu_a + \sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} \right\} \\
& \cup \left\{ \begin{aligned} & \forall a \notin I^*, \hat{\mu}_a(t-1) - \sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} < \mu_a \\ & u_{I(t)}(t) > u_{I^*}(t) : \\ & \forall a \in I^*, \hat{\mu}_a(t-1) + \sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} > \mu_a \end{aligned} \right\}
\end{aligned}$$

First we prove an upper bound about the first event.

$$\begin{aligned}
& \sum_{t=1}^T \Pr \left\{ \exists a \in I^*, \hat{\mu}_a(t-1) \leq \mu_a - \sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} \right\} \\
& \leq \sum_{t=1}^T \sum_{a \in I^*} \Pr \left\{ \hat{\mu}_a(t-1) \leq \mu_a - \sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} \right\} \\
& = \sum_{t=1}^T \sum_{a \in I^*} \sum_{n=1}^{t-1} \Pr \left\{ \hat{\mu}_{a,n} \leq \mu_a - \sqrt{\frac{\eta \ln(t)}{2n}}, N_a(t-1) = n \right\} \\
& \leq \sum_{t=1}^T \sum_{a \in I^*} \sum_{n=1}^{t-1} \Pr \left\{ \hat{\mu}_{a,n} \leq \mu_a - \sqrt{\frac{\eta \ln(t)}{2n}} \right\}
\end{aligned}$$

By using Hoeffding's inequality (Fact 1), we have

$$\Pr \left\{ \hat{\mu}_{a,n} \leq \mu_a - \sqrt{\frac{\eta \ln(t)}{2n}} \right\} \leq t^{-\eta}.$$

Then we can find that

$$\begin{aligned}
\sum_{t=1}^T \sum_{a \in I^*} \sum_{n=1}^{t-1} \Pr \left\{ \hat{\mu}_{a,n} \leq \mu_a - \sqrt{\frac{\eta \ln(t)}{2n}} \right\} & \leq \sum_{t=1}^T \sum_{a \in I^*} \sum_{n=1}^{t-1} t^{-\eta} \\
& \leq \sum_{t=1}^T \sum_{a \in I^*} t^{1-\eta} \\
& = M \sum_{t=1}^T t^{1-\eta}.
\end{aligned}$$

This can be upper bounded by indefinite integral. If $\eta > 2$, we find that

$$\sum_{t=1}^T t^{1-\eta} \leq \int_0^\infty t^{1-\eta} dt = -\frac{1}{2-\eta} = \frac{1}{\eta-2}.$$

In summary,

$$\sum_{t=1}^T \Pr \left\{ \exists a \in I^*, \hat{\mu}_a(t-1) \leq \mu_a - \sqrt{\frac{\eta \ln(t)}{N_a(t-1)}} \right\} \leq \frac{M}{\eta-2}$$

if $\eta > 2$.

The same procedure will leads

$$\sum_{t=1}^T \Pr \left\{ \exists a \notin I^*, \hat{\mu}_a(t-1) \geq \mu_a + \sqrt{\frac{\eta \ln(t)}{N_a(t-1)}} \right\} \leq \frac{M(N-1)}{\eta-2}.$$

if $\eta > 2$.

The upper bound on third event

$$\begin{aligned} & \left\{ \sum_{a \in I(t) \setminus I^*} \left(\hat{\mu}_a(t-1) + \sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} \right) \right. \\ & \quad \left. > \sum_{a \in I^* \setminus I(t)} \left(\hat{\mu}_a(t-1) + \sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} \right) \right\} \\ & \subseteq \left\{ \sum_{a \in I(t) \setminus I^*} \left(\mu_a + 2\sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} \right) > \sum_{a \in I^* \setminus I(t)} \mu_a \right\} \\ & = \left\{ \sum_{a \in I(t) \setminus I^*} 2\sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} > \Delta_{I(t)} \right\}. \end{aligned}$$

This never happens if

$$\sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} < \frac{\Delta_{I(t)}}{M\sqrt{2\eta \ln(T)}}$$

since

$$\left\{ \sum_{a \in I(t) \setminus I^*} 2\sqrt{\frac{\eta \ln(t)}{2N_a(t-1)}} > \Delta_{I(t)} \right\} \subseteq \left\{ \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \frac{\Delta_{I(t)}}{M\sqrt{2\eta \ln(T)}} \right\}$$

holds. Therefore, we can find the upper bound about third term by

$$\sum_{t=1}^T \Delta_{I(t)} \mathbb{I} \left\{ \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} < \frac{\Delta_{I(t)}}{M\sqrt{2\eta \ln(T)}} \right\}.$$

This can be upper bounded as

$$\sum_{t=1}^T \Delta_{I(t)} \mathbb{I} \left\{ \sum_{a \in I(t)} \frac{1}{\sqrt{N_a(t-1)}} > \frac{\Delta_{I(t)}}{M\sqrt{2\eta \ln(T)}} \right\} \leq \frac{32\eta M^{7/3} N}{\Delta_{\min}} \ln(T)$$

by Lemma 8.

Finally, putting everything together, we get

$$R_{\pi}(T) \leq \frac{32\eta M^{7/3} N}{\Delta_{\min}} \ln(T) + \frac{MN\Delta_{\max}}{\eta - 2},$$

which concludes the proof. $\square$

4.4 Summary

We improved the previous best upper bound on the expected regret of LLR by a factor of $\Theta(M^{2/3})$ for the matching-selection bandit problem on the complete bipartite graph $K_{M,N}$. Although this bound is still larger than the bound proved for the optimally tuned ϵ -greedy policy by a factor of $(M^{1/3})$, that is smallest upper bound among anytime distribution-free policies for matching-selection problem.

Chapter 5

Conclusion

In this paper, we consider two kinds of multi-armed bandit problem.

First one is a budgeted multi-armed bandit problem. For this problem, we designed two policies KL-UCB-SC and UCB-SC. As for KL-UCB-SC, we proved that it is an asymptotically optimal policy for a budgeted multi-armed bandit problem. Note that KL-UCB-SC is the first policy which has been proved to be asymptotic optimal. But KL-UCB-SC suffered its prohibitive computation time. In order to overcome this drawback, we also designed UCB-SC. In our computational experiments, KL-UCB-SC+ worked almost the best among the known policies in terms of average regret and UCB-SC+ worked the best among closed-form-based policies, such as UCB-BV1 and PD-BwK.

As noticed before, in order to calculate an action index for KL-UCB-SC, we had to solve an optimization problem, which takes more time than existing policies. An efficient way to compute the index would be an interesting future work. Although the known upper bound of regret was too large, a policy BTS, based on Thompson Sampling worked unexpectedly well in our experiments. Given this, we conjectured that BTS is asymptotically optimal.

Second one is a matching-selection bandit problem, which can be considered as a special instance of combinatorial semi-bandit problem in the case with the matching-selection bandit problem using a new technique. For this problem, we give a refined analysis of a policy LLR, which is for general combinatorial semi-bandit problem. As a result, for a matching-selection bandit problem on complete bipartite graph $K_{M,N}$, we improved an upper bound of expected regret by a factor of $\Theta(M^{2/3})$.

Bibliography

- [1] W. R. Thompson, On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, **25**-3(1933), 285–294.
- [2] H. Robbins, Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, **58**-5(1952), 527–535.
- [3] P. Auer, N. Cesa-Bianchi and P. Fischer, Finite-time analysis of the multiarmed bandit problem. *Machine learning*, **47**-2(2002), 235–256.
- [4] P. Auer, N. Cesa-Bianchi, Y. Freund and R. E. Schapire, Gambling in a Rigged Casino: The Adversarial Multi-armed Bandit Problem. *Proceedings of 36th Annual Symposium on Foundations of Computer Science*, 1995, 322–331.
- [5] P. Auer, N. Cesa-Bianchi, Y. Freund and E. Schapire, Robert, The non-stochastic multiarmed bandit problem. *SIAM Journal on Computing*, **32**-1(2003), 48–77.
- [6] J.-Y. Audibert, S. Bubeck and R. Munos, Best Arm Identification in Multi-Armed Bandits. *Proceedings of the 23rd Conference on Learning Theory*, 2010, 41–53.
- [7] A. Garivier and O. Cappé, The KL-UCB algorithm for bounded stochastic bandits and beyond. *Proceedings of the 22nd International Conference on Algorithmic Learning Theory*, 2011, 359–376.
- [8] O. Cappé, A. Garivier, O.-A. Malliard, R. Munos and G. Stoltz, Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation. *Annals of Statistics*, **41**-3(2013), 1516–1541.
- [9] J. Honda and A. Takemura, An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. *Proceedings of the 23rd Annual Conference on Learning Theory*, 2010, 67–79.

- [10] J. Honda and A. Takemura, An Asymptotically Optimal Policy for Finite Support Models in Multiarmed Bandit Problem. *Machine Learning*, **85**(2011), 361–391.
- [11] E. Kaufmann, N. Korda and R. Munos, Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, 2012, 199–213.
- [12] T. L. Lai and H. Robbins, Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, **6-1**(1985), 4–22.
- [13] W. Ding, T. Qin, X.-D. Zhang and T.-Y. Liu, Multi-armed Bandit with Budget Constraint and Variable Costs. *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013, 232–238.
- [14] Y. Gai, B. Krishnamachari and R. Jain, Learning multiuser channel allocations in cognitive radio networks: a combinatorial multi-armed bandit formulation. *IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, 2010, 571–579.
- [15] L. Tran-Thanh, A. Chapman, E. M. d Cote, A. Rogers and N. R. Jennings, ϵ -First Policies for Budget-Limited Multi-Armed Bandits. *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010, 1211–1216.
- [16] L. Tran-Thanh, A. Rogers, A. Chapman and N. R. Jennings, Knapsack Based Optimal Policies for Budget-Limited Multi-armed Bandits. *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012, 1134–1140.
- [17] Y. Xia, H. Li, T. Qin, N. Yu and T.-Y. Liu, Thompson Sampling for Budgeted Multi-armed Bandits. *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015, 3960–3966.
- [18] E. Kaufmann, Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources. Ph. D. thesis of Telecom ParisTech, 2014.
- [19] V. Anantharam, P. Varaiya and J. Walrand, Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards. *Transactions on Automatic Control*, **32-11**(1987), 968–976.

- [20] Y. Gai, B. Krishnamachari and R. Jain, Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, **20**-5(2012), 1466–1478.
- [21] W. Chen, Y. Wang and Y. Yuan, Combinatorial Multi-Armed Bandit: General Framework, Results and Applications. *The Proceedings of the 30th International Conference on Machine Learning*, 2013, 151–159.
- [22] N. Cesa-bianchi and G. Lugosi, Combinatorial bandits. *Journal of Computer and System Sciences*, **78**(2012), 1404–1422.
- [23] M. S. Talebi and A. Proutiere, An Optimal Algorithm for Stochastic Matroid Bandit Optimization. *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*, 2016, 548–556.
- [24] J. Komiyama, J. Honda and H. Nakagawa, Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays. *The Proceedings of the 32nd International Conference on Machine Learning*, 2015, 1152–1161.
- [25] T. Uchiya, A. Nakamura and M. Kudo, Algorithms for Adversarial Bandit Problems with Multiple Plays. *Proceedings of the 21st International Conference on Algorithmic Learning Theory*, 2010, 375–389.
- [26] W. Hoeffding, Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, **58**-301(1963), 13–30.
- [27] S. Bubeck, Jeux De Bandits Et Foundations Du Clustering. Ph. D. thesis of Université Lille 1, 2010.
- [28] A. Badanidiyuru, R. Kleinberg and A. Slivkins, Bandits with Knapsacks. *Proceedings of the 54th IEEE Annual Symposium on Foundation of Computer Science*, 2013, 207–216.
- [29] H. W. Kuhn, The Hungarian method for the assignment problem. *Naval research logistics quarterly*, **2**-1(1955), 83–97.
- [30] D. P. Helmbold and M. K. Warmuth, Learning permutations with exponential weights. *The Journal of Machine Learning Research*, **10**(2009), 1705–1736.

- [31] K. Mohan and O. Dekel, Online bipartite matching with partially-bandit feedback. *NIPS Workshop on Discrete Optimization in Machine Learning*, 2011.
- [32] A. György, T. Linder, G. Lugosi and G. Ottucsák, The On-Line Shortest Path Problem Under Partial Monitoring. *Journal of Machine Learning Research*, **8**(2007), 2369–2403.
- [33] J.-Y. Audibert, S. Bubeck and G. Lugosi, Minimax Policies for Combinatorial Prediction Games. *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, 107–132.
- [34] G. Neu and G. Bartók, An Efficient Algorithm for Learning with Semi-Bandit Feedback. *Proceedings of 24th International Conference on Algorithmic Learning Theory*, 2013, 234–248.

謝辞

本研究に取り組むにあたり，指導教官である北海道大学大学院 情報科学研究科 情報理工学専攻 数理科学講座 情報認識学研究室の 中村 篤祥 准教授には，長年にわたり専門的な知識や研究の進め方など多くの貴重な教示を賜り，深く御礼申し上げます。

また，所属研究室の長であります 工藤 峰一 教授には貴重な御指導を頂きましたことを御礼申し上げます。

さらに，同研究室の先輩方にはご多忙の中研究や発表にとどまらず，様々な御協力・御指導を頂けたことを感謝します。

研究室外につきましても，各種研究会やワークショップでの議論は大変助けになりました。それらで議論に参加してくださった研究者の方々にも御礼申し上げます。その中でも特に東京大学 小宮山 純平 さんにつきましては，学位取得にあたり執筆した主要な論文に共著者としてご協力いただいたこともあり，特段に感謝申し上げます。

博士課程学生として過ごした4年間のうち，後半2年間の間には2つの研究プロジェクトにリサーチ・アシスタントとして関わってきました。

ひとつは 科研費・基盤 (S)「離散構造処理系の基盤アルゴリズムの研究」で，こちらのプロジェクトでは代表者である 湊 真一 教授のみならず，石畠 正和 特任助教及び他のリサーチアシスタントの面々と大変有意義な時間を過ごすことができました。

もうひとつは 小松崎 民樹 教授を代表者とした JST/CREST [情報計測]「一細胞ラマン計測と情報科学の融合による細胞診断の迅速解析技術の開発」で，医学や光学といった分野の研究者を交えた議論ができたことは，今後の研究者としてのキャリア形成の助けになることと思います。

最後になってしまいましたが，高専を卒業してから更に8年もの間，働かずに過ごすことを認め，そして支えてくれた両親に感謝します。