Asymptotically Optimal Multi-armed Bandit Algorithms Aimed at Online Contents Selection
(オンラインコンテンツ選択のための漸近最適な多腕バンディットアルゴリズム)

Junpei Komiyama (小宮山 純平)

Abstract

A multi-armed bandit problem is a crystallized instance of a sequential decision-making problem in an uncertain environment. The history of this problem at least goes back to the 1930s, and it has recently attracted attention in the machine learning community. This problem involves conceptual entities called arms, and a forecaster who tries to identify good arms from bad ones. At each round, the forecaster draws one of the K arms and receives a corresponding reward. The aim of the forecaster is to maximize the cumulative reward over rounds, which is achieved by running an algorithm that balances the exploration (acquisition of information) and the exploitation (utilization of information). The notion of strong consistency introduced by Lai and Robbins (1985) defines the optimal balance between the exploration and the exploitation for a certain class of robust algorithms. At the beginning of the 2000s, an algorithm based on the upper confidence bound was established. Around that time, people found that many problems in web systems that involve uncertainty are related to the multi-armed bandit problem.

In each round of the multi-armed bandit problem, the algorithm selects an arm and receives a reward. In other words, the three core notions in the multi-armed bandit problem are (i) the sequential selection of arms, (ii) the criterion of selection (i.e., single arm selecting), and (iii) the reward feedback. However, these three notions are usually violated in practical applications: there are gaps between the multi-armed bandit framework and practical problems in web systems to which we would like to apply it.

In this thesis, we first review the history and the state-of-the-art framework of the multi-armed bandit. Then, we propose three extensions to the multi-armed bandit problem, which are intended to fill these gaps. Namely, (i) we propose the lock-up bandit problem, which models technical restrictions on the sequential selection of arms. (ii) Motivated by the problem of online advertisement placement, we study a multiple-play version of the multi-armed bandit problem. We propose an extension of the Thompson sampling algorithm and show its effectiveness both theoretically and empirically. Moreover, (iii) motivated by problems arising in the information retrieval domain, we study the dueling bandit problem, a variant of the multi-armed bandit problem in which only the result of a pairwise comparison is available. A family of algorithms based on the likelihood function is proposed, and their effectiveness is verified.

Contents

Chapter 1

Introduction

A multi-armed bandit problem is a crystallized instance of a sequential decision-making problem in an uncertain environment, and it can model many real-world scenarios. This problem involves conceptual entities called arms, and a forecaster who tries to identify good arms from bad ones. At each round, the forecaster draws one of the arms and receives a corresponding reward. The aim of the forecaster is to maximize the cumulative reward over rounds, which is achieved by running an arm selection algorithm that balances exploration (acquisition of information) and exploitation (utilization of information). Assuming that the rewards of each arm are sampled from a distribution that does not change over rounds, the maximization of the cumulative reward boils down to choosing the distribution giving the largest expectation for the most rounds.

Although the exact origin of the multi-armed bandit problem is not clear, the essential idea behind it first arose in very old papers. Motivated by sequential experimental design, Thompson [1933] studied a method to compute the probability that one arm is superior to another, which can be considered a two-armed bandit problem. Robbins [1952] also studied the two-armed bandit problem[*1]. He showed that, from the strong law of large numbers, the reward per round of the forecaster can be made arbitrarily close to the best possible one.

The exact maximization of the cumulative reward is very hard to achieve, and the introduction of a discount factor made the problem much easier. Bellman [1956] devised an algorithm that maximizes the cumulative reward for some class of Bayesian discounted two-armed bandit problems. This result was generalized by the subsequent line of research: arguably, the most seminal result of this formulation is the so-called Gittins index [Gittins and Jones, 1974], which by computing for each arm gives an algorithm that maximizes the cumulative discounted reward.

In contrast, the maximization of the undiscounted cumulative reward eluded understanding for a long time, due in part by the lack of convergence. A breakthrough occurred in two papers in the 1980s, when Lai and Robbins [1985] derived the asymptotic possible

[*1] Note that Robbins did not cite Thompson's work.

best performance of an algorithm without prior knowledge of the model parameter. Lai [1987] produced a similar result in view of Bayesian statistics. They also proposed upper confidence bound algorithms that are optimal in a sense of these performance metrics. The idea of an upper confidence bound was simplified later in Agrawal [1995b]. These papers established the basic framework of the stochastic bandit, which is commonly used in today's machine-learning community. Auer et al. [2002a] proposed the UCB1 algorithm, which is very simple yet has strong theoretical properties. Thanks to its simplicity, UCB1 has been widely used in the machine learning community, and many extensions to it have been proposed.

Several years before the appearance of UCB1, Abe and Nakamura [1999] applied the bandit framework to the problem of optimal placement of online advertisements. Around that time, the multi-armed bandit started to attract attention from the machine learning community. People found that the framework of the bandit problem can be generalized to a wider class of problems. Although the basic problem is about identifying the optimal arm (i.e., the arm with the largest expected reward) among several, in the most general sense, the bandit problem is an important subclass of sequential learning where the available feedback is limited. In particular, a wide variety of web related problems that involve uncertainty can be modeled as an extension of a bandit problem. To see this, take the example of online advertising. Consider a website that has a pool of relevant advertisements (ads). There is space on the website to show an ad. When a user arrives, the website selects an ad from the pool. If the user is interested in the ad, he/she clicks it. A significant fraction of online advertising uses the pay-per-click model, and as such maximization of revenue is equivalent to maximization of the number of clicks. The website does not know the percentage of users clicking on each ad (click-through rate). This ad selection problem is a representative example of the multi-armed bandit problem. Namely, the ads, the users, and the clicks correspond to the arms, the rounds, and the rewards of the problem. Feedback is only available from the displayed ad, and feedback is never available from the undisplayed ads: the feedback is limited in this sense. The multi-armed bandit problem has even wider application to web systems: examples include not only online advertising, but also content optimization [Agarwal et al., 2009, Scott, 2010], search engine optimization [Radlinski et al., 2008a], network routing [Awerbuch and Kleinberg, 2004], and recommender systems [Li et al., 2010].

1.1 Balancing Exploration and Exploitation

The central theme of the multi-armed bandit problem is finding the optimal balance between the exploration and the exploitation of information. Regret, which is defined as the difference between the cumulative rewards of the optimal arm and the algorithm, corresponds to how much information is required for learning the structure of the problem.

Therefore, an algorithm minimizing the regret is the one that fully utilizes the feedback. The theory of strongly consistent algorithms, which we describe later, is simple enough to understand intuitively and makes sense when taking the uncertainty of the parameters into consideration. This theory defines an amount of regret required by bandit algorithms in the asymptotical sense, which is a reasonable answer as to the optimal balance between the exploration and the exploitation. We argue that most of the existing bandit algorithms are not asymptotically optimal in the sense of regret. In applying the bandit framework, one usually resorts to well-known variants of algorithms, such as ϵ-greedy and UCB1. However, these simple algorithms sacrifice the performance by choosing a conservative confidence bound for the sake of ease of analysis. To design an asymptotically optimal algorithm, the number of draws of each arm must be controlled, and a refined analysis of the draws is required. A few algorithms, such as KL-UCB, Thompson sampling, and Deterministic Minimum Empirical Divergence, are known to be asymptotically optimal. We show a general idea that is common to all of these optimal algorithms.

1.2 Rethinking the Bandit Framework

The framework of the multi-armed bandit is quite simple: yet it is powerful enough to model many situations that involve sequential decision-making. In particular, we focus on applying the multi-armed bandit framework to content optimization problems on web systems. Although the framework is flexible, in many situations, there are non-trivial gaps between it and the content optimization. The contribution of this thesis is to propose a way to fill these gaps by extending the stochastic bandit framework.

In each round of the multi-armed bandit problem, the algorithm selects an arm and receives a reward. In other words, the three core notions in the multi-armed bandit problem are (i) the sequential selection of arms, (ii) the criterion of selection (i.e., single arm selecting), and (iii) the reward feedback. In this thesis, we rethink them. Figure 1.1 describes the stochastic bandit problem and the three extensions that are studied in this thesis. A natural question is whether the exploration and the exploitation can be balanced in these extensions: we give a positive answer to this question by proposing algorithms that minimize the regret in these extensions.

1.2.1 A lock-up restriction for the multi-armed Bandit Problem

The multi-armed bandit framework assumes that at each round an algorithm can select an arm freely. When we deploy bandit algorithms in web systems, this assumption is usually violated by technological limitations.

The most well-studied restriction on the bandit problem is the one called the switching cost [Jun, 2004, Mahajan and Teneketzis, 2008]: it penalizes switching between one arm and another. The switching cost is mainly motivated by economic considerations: it is

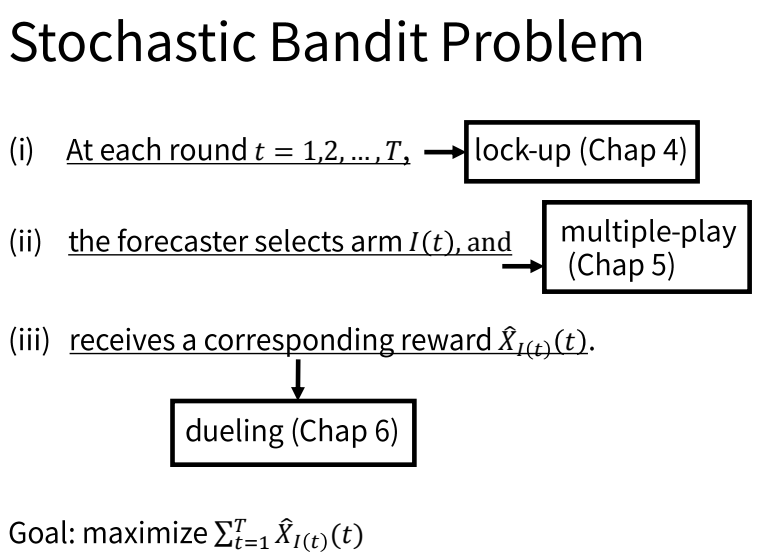


Fig. 1.1. The standard stochastic bandit problem and its extensions that are studied in this thesis.

often the case that changing an option is more costly than staying with it. The switching cost is also motivated by experimental design: it models the cost of switching between experimental settings. Although the algorithm is motivated to stay with the current arm, it may change the arm at any round by paying the cost: in this sense, the switching cost is a soft constraint. While the switching cost can model these problems, it cannot model the technical limitations that arise in web systems. For instance, in the case of online advertisements, updates of the algorithm parameters are delayed by the latency of the system and the users' feedback. These delays are not reduced by paying switching costs, and they are better modeled by a harder constraint.

Taking this into consideration, we propose a version of the bandit problem called lock-up bandits. This problem involves lock-up periods, during which an algorithm cannot switch the arm. The lock-up period restriction is more stringent than the switching cost in the sense that the algorithm cannot change the arm during a period. Existing bandit algorithms, such as UCB1 and ϵ-greedy, cannot be directly applied to a problem with lock-up periods because they choose an arm in each round without considering the lock-up constraint. We derive a meta-algorithm that converts certain bandit algorithms, including UCB1 and ϵ-greedy, into ones that are compatible with lock-up periods. The performance of these algorithms is verified theoretically and empirically: the price to pay for the lock-up periods is linear to the longest lock-up period. Interestingly, the derived regret bound does not depend on the number of the lock-up periods. The results of computer simulations support this conclusion.

1.2.2 On multiple plays of the multi-armed bandit problem

In some applications such as online advertising, multiple entities that correspond to arms are recommended to a user: a website usually has several slots for ads, and thus multiple ads are selected at each round. In the standard multi-armed bandit problem, a forecaster at each round selects single arm, and it cannot directly model a multiple ad selection problem. However, it is possible to model this multiple ad selection as a bandit problem by considering a set of ads as a single arm: let L be the number of the slots in which the ads are placed. The number of L subsets of K ads is ${}_KC_L$, which is proportional to K^L for $L \ll K$: modeling each L-subset as an arm is prohibitive in the case $L > 1$. Instead, we consider an extension of a multi-armed bandit problem in which L arms among K are chosen at each round. Several algorithms that are asymptotically optimal in the standard single-play bandit problem, such as KL-UCB [Lai, 1987, Cappé et al., 2013], DMED [Honda and Takemura, 2010], and Thompson sampling [Thompson, 1933], are known. At a glance, it appears obvious that these algorithms are asymptotically optimal for the multiple-play case. Interestingly, it turns out to be highly non-trivial to prove their optimality, even the simplest case in which there is no correlation between the click-through rates of ads; to achieve an asymptotically optimal regret in the multiple-play bandit problem, an algorithm must have plausible combinatorial properties. We show that the multiple-play extension of Thompson sampling has such a property.

1.2.3 On pairwise comparison in the multi-armed Bandit Problem: the Dueling Bandit Problem

The reward in the multi-armed bandit problem is absolute metric in the sense that a higher reward is better. However, the availability of such an absolute metric is not always the case, especially the problems involving the preferences of humans. Even if absolute feedback is available, it may be biased. The most well-studied case is the ranking function evaluation used in information retrieval. Regarding the evaluation of ranking functions based on users' implicit feedback, absolute metrics (e.g., clicks per query) do not reflect the retrieval quality of the ranking functions, whereas pairwise comparison based on the interleaved ranking function [Joachims, 2003] provides a more reliable metric [Radlinski et al., 2008b]. Motivated by such a search engine ranker evaluation problem, Yue et al. [2009] proposed the dueling bandits problem. In this problem, an algorithm selects a pair of arms at each round and receives information that indicates which of the two arms is preferred. There are several criteria about the best arm based on the relative comparison. In particular, we study the problem of finding the arm that satisfies the Condorcet winner criterion [Urvoy et al., 2013]. Here, we are interested in the exploration and exploitation

trade-off in pairwise comparison: to how many users can we recommend an optimal arm throughout the rounds?

1.3 Study of Bandit Algorithms: Theory Meets Practice

In the study of algorithms, one often observes the gap between theory and practice: For instance, it is often the case that an algorithm with a good theoretical property (e.g., a performance guarantee) does not perform well in practical situations. Moreover, it is often the case that such a theoretical algorithm is quite involved; it may contain many artificial terms for ease of analysis. In practical situations, one usually resorts to more practical algorithms that have weak (or sometimes no) theoretical guarantees. In my experience, except for experts on algorithms, users prefer simple and practical algorithms over complex ones. Therefore, the theoretical study of algorithms is somewhat different from work aimed at making practically useful ones. Fortunately, this is not the case with the bandit algorithms in this thesis: they are examples in which theory benefits practice.

1.4 Structure of This Thesis

The rest of this thesis is organized as follows. In Chapter 2, we introduce the standard multi-armed bandit problem. Historically speaking, there are three approaches to setting the rewards, and we compare them. Furthermore, among them, the stochastic approach is currently the most widely used in the machine learning community. We explain that strong consistency is an essential component of the stochastic bandit problem. An asymptotical regret lower bound of a strongly consistent bandit algorithm is derived.

Henceforth, we adopt the stochastic approach. In Chapter 3, we introduce the standard stochastic bandit algorithms. We explain the general idea that is common to efficient stochastic bandit algorithms: they control the number of draws on suboptimal arms so that its regret asymptotically matches the regret lower bound . Understanding this idea is crucial when one tries to extend multi-armed bandit algorithms. Moreover, we show the results of simulations illustrating the performance of these algorithms in practical situations.

The following three chapters discuss extensions of the bandit problem. In Chapter 4, we propose the lock-up bandit problem. We are interested in how the restriction on selection affects the exploration and exploitation balance in the multi-armed bandit problem.

In Chapter 5, we study the multiple-play multi-armed bandit problem, where $L \geq 1$ arms are selected in each round. In the case of multiple selection, drawing several suboptimal arms decreases the reward: we show that this problem can be circumvented by using an extension of the Thompson sampling algorithm. The stochastic nature of the algorithm is used to prove the asymptotic optimality of the algorithm.

In Chapter 6, we study the dueling bandit problem. Under the Condorcet assumption [Urvoy et al., 2013] on the preferences of the arms, we discuss the optimal balance between exploration and exploitation. We derive an asymptotical regret lower bound and devise an algorithm whose performance asymptotically matches the regret lower bound.

In Chapter 7, we give concluding remarks and discuss some of the existing studies on the extension of the multi-armed bandit problem. The bandit problem is extensively studied in the literature: we mainly restrict our interest to the fields of machine learning and data mining.

Chapter 2

Framework of Multi-armed Bandit Problem

In this chapter, we discuss the framework of the standard multi-armed bandit problem. Historically speaking, there are three major approaches to the bandit problem: namely, the Bayesian, stochastic, and adversarial. We will explain these approaches and discuss their advantages and disadvantages. The notation used in this chapter is summarized in Table 2.1.

2.1 Multi-armed Bandit Problem

In this section, we explain the basic multi-armed bandit framework. Let there be K arms. The notation "$A := B$" means that "A equals B by definition". At each round $t = 1, 2, \ldots, T$, the forecaster selects an arm $I(t) \in [K] := \{1, 2, \ldots, K\}$, then receives a corresponding reward $\widehat{X}_{I(t)}(t)$. The forecaster's objective is to maximize the sum of rewards. To do so, the forecaster should use an algorithm that selects the arm with the largest expected reward, which we call the optimal arm, for most rounds. Much effort has been devoted to finding efficient algorithms for solving the multi-armed bandit problem.

Regarding the reward model, three approaches have been considered in the literature: Bayesian, stochastic, and adversarial (Figure 2.1).

- **Bayesian approach:** typically, each arm is associated with a distribution that belongs to some parameterized family of functions (e.g., Bernoulli or Normal). The reward at each round is an i.i.d. sample from the distribution associated with the selected arm. This approach adopts a Bayesian view on the distribution from which the rewards are generated. The objective of the forecaster is to maximize the discounted sum of rewards in the Bayesian sense. Most results involve a discount factor: an algorithm weight the current reward more than the future rewards. The index theorem formalized by Gittins [Gittins and Jones, 1974] gives a construction

Table. 2.1. Notation used in Chapter 2.

$\mathbb{R}^+$	:=	$(0, +\infty)$.
$\mathbf{1}\{A\}$	:=	1 if A is true and 0 otherwise.
K	:=	Number of the arms.
$[K]$	:=	$\{1, 2, \ldots, K\}$.
T	:=	Number of the rounds.
$I(t)$	:=	The arm that is selected in round t.
$\theta = (\theta_1, \ldots, \theta_K)$	:=	Model parameters.
P_{θ_i}	:=	Probability distribution from which the reward of arm i is generated.
$\widehat{X}_i(t)$	:=	Reward of arm i at round t.
μ_i	:=	Mean reward of arm i.
Δ_i	:=	$\mu_1 - \mu_i$.
$\widehat{\mu}_i(t)$	:=	Empirical mean reward of arm i at round t.
$N_i(t)$	:=	Number of rounds in which arm i is selected before round t: that is, $\sum_{t'=1}^{t-1} \mathbf{1}\{I(t') = i\}$.
$d(p, q)$	:=	The KL divergence between distributions with parameters p and q. In the case of Bernoulli distributions, $d(p, q) = p \log(p/q) + (1-p) \log((1-p)/(1-q))$.

of the optimal algorithm. This approach is suitable for the case in which prior knowledge about the arms is available.

- **Stochastic approach:** like the Bayesian approach, each arm is associated a parameterized distribution, and the reward at each round is an i.i.d. sample from the distribution associated with the selected arm. This approach seeks an algorithm that performs well with any set of parameters; in the sense that it does not require the prior knowledge of the parameters, this approach is inherently robust. The objective is to maximize the undiscounted sum of rewards.
- **Adversarial approach:** in this approach, no assumption on the rewards is imposed: yet there are some ingenious randomized algorithms that perform effectively. The objective is to maximize the undiscounted sum of rewards. This approach is suitable for cases where it is difficult to model how the rewards are distributed.

Model Parameter:	Model Parameter:	Model Parameter:
Given (prior)	Unknown	Nonparametric
Objective:	Objective:	Objective:
Bayes reward maximization.	Regret minimization	Regret minimization
Future rewards:	Future rewards:	Future rewards:
Discounted	Undiscounted	Undiscounted
(a) Bayesian	(b) Stochastic	(c) Adversarial

Fig. 2.1. Approaches for solving the multi-armed bandit problem

2.2 Bayesian Approach

In the Bayesian approach (Figure 2.2), the reward of each arm i is drawn from a family of distributions P_{θ_i}, which is parameterized by θ_i. The model parameters of the distributions $\theta = (\theta_1, \ldots, \theta_K) \in \Theta$ is drawn from a known prior Π. Let $\widehat{X}_i(t) \sim P_{\theta_i}$ be the reward of arm i and $\mu_i = \mathbb{E}[\widehat{X}_i]$ be the expected reward of arm i. Let $\beta \in (0, 1)$ be a discount factor. The goal of the forecaster is to maximize the discounted cumulative reward:

$$\mathbb{E}_{\Pi}\left[\sum_{t=1}^{T} \beta^{t-1} \widehat{X}_{I(t)}(t)\right], \tag{2.1}$$

where the expectation is taken over the prior. The introduction of the discount factor makes the analysis easier since the discounted sum of rewards converges. The most seminal result in this setting is the Gittins index theory: each arm i is characterized by an index function $G_i = G_i(t)$. Gittins and Jones [1974] showed that selecting the arm of the largest index $i^* = \arg\max_{i \in [K]} G_i$ maximizes the discounted cumulative reward of the inequality (2.1). Note that, the index theorem applies to a more general case of K bandit processes in which each arm has its state that evolves for each draw. For the ease of discussion, we do not go into the details about the bandit processes.

The introduction of the discount factor implies that the future reward is less important than the current one, and, as a result, it encourages more exploitation in early rounds. The smaller discounted factor is, the more short-term revenue the algorithm seeks. In the case of web systems, the rounds correspond to each user who accesses the website. The optimal choice of the discount factor depends on how many users the website can expect to arrive, which sometimes is difficult to estimate beforehand.

A set of model parameters $\theta = (\theta_1, \ldots, \theta_K) \in \Theta$ is drawn from the prior distribution Π.
Input: K (number of arms), β (discounted factor)
At each round $t = 1, ..., T$, the algorithm

1. selects an arm I_t, and
2. receives a reward $\widehat{X}_{I(t)}(t) \sim P_{\theta_{I(t)}}$.

Goal: maximize the cumulative reward $\sum_{t=1}^{T} \beta^{t-1} \widehat{X}_{I(t)}(t)$.

Fig. 2.2. The Bayesian bandit problem

2.3 Stochastic Approach

In the stochastic approach (Figure 2.3), the reward $\widehat{X}_i(t)$ from arm i is an i.i.d. sample from a distribution P_{θ_i}. The objective of the forecaster is to maximize the cumulative reward:

$$\mathbb{E}\left[\sum_{t=1}^{T} \widehat{X}_{I(t)}(t)\right]. \tag{2.2}$$

Unlike the Bayesian approach, the reward in (2.2) is undiscounted: it equally weights the current reward and the rewards in the future rounds.

2.3.1 Optimal arm and regret

Let $\mu_i = \mathbb{E}[\widehat{X}_i(t)]$ be the expected reward of arm i. Without loss of generality, we can assume $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. For ease of discussion, let the arm of the largest reward be unique[*1]: that is, $\mu_1 > \mu_2$. If all the model parameters are known, the best option is to select the arm with the largest mean at each round. In this sense, we call arm 1 the optimal arm and the others suboptimal arms. Since the algorithm does not know the parameters, it needs to acquire information from all arms (exploration). Meanwhile, it should use the current information to obtain a better short-term reward (exploration). A good algorithm should balance exploration and exploitation, which is measured in terms of the regret.

Let $\Delta_i = \mu_1 - \mu_i$. The regret, which is the difference between the rewards of the optimal arm and the algorithm, is defined as

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} (\mu_1 - \mu_{I(t)}) = \Delta_{I(t)}.$$

[*1] This assumption can be removed: for example, see Appendix A in Agrawal and Goyal [2012].

Model Parameters $\theta = (\theta_1, \ldots, \theta_K) \in \Theta$ (unknown)
Input: K (number of arms)
At each round $t = 1, ..., T$, the algorithm

1. selects an arm I_t, and
2. receives a reward $\widehat{X}_{I(t)}(t) \sim P_{\theta_{I(t)}}$.

Goal: maximize the cumulative reward $\sum_{t=1}^{T} \widehat{X}_{I(t)}(t)$.

Fig. 2.3. The stochastic bandit problem

Note that $\Delta_i > 0$ for all $i \neq 1$ and $\Delta_1 = 0$: the regret is non-negative, and increases by Δ_i if we select arm i once. Clearly, maximization of the cumulative reward is equivalent to minimization of the regret. The expectation of the regret $\mathbb{E}[\mathrm{Reg}(T)]$ measures the performance of an algorithm. The advantage of using regret is that it clarifies the amount of exploration and enables us to discuss the performance of the algorithms in terms of exploration and exploitation. Robbins [1952] is one of the first papers that derived a non-trivial result on the undiscounted cumulative reward. They studied a two-armed bandit problem and showed that it is possible to construct an algorithm with its regret per round approaches zero: namely, there exists an algorithm such that

$$\lim_{T \to +\infty} \mathbb{E}\left[\frac{\mathrm{Reg}(T)}{T}\right] \to 0.$$

2.3.2 Strong consistency and regret lower bound

In Section 2.3.1, we defined the regret, which by using an ingenious algorithm can be sublinear to the number of rounds T. A natural question is how fast the regret per round approaches zero. Given an algorithm with certain properties, a logarithmic regret can be shown to be the best possible performance. In this subsection, we formalize this result.

An algorithm is strongly consistent if

$$\mathbb{E}[\mathrm{Reg}(T)] = o(T^a)$$

for any $a > 0$ and any set of model parameters $\theta \in \Theta$. At In a word, a strongly consistent algorithm is "uniformly good" over all instances. Let arm $i \neq 1$ be arbitrary and $N_i(t)$ be the number of rounds before t in which arm i is selected. Lai and Robbins [1985] showed that, there exists an asymptotic lower bound on $N_i(T)$ for any strongly consistent algorithm. Let $d(\theta_i, \theta_1)$ be the Kullback-Leibler (KL) divergence between two distributions with their parameters θ_i and θ_1. For any bandit problem with itswhose rewards are drawn from any single parameter exponential family of distributions, the following inequality

holds:

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\log T} \geq \frac{1}{d(\theta_i, \theta_1)}. \tag{2.3}$$

An asymptotic regret lower bound easily follows from (2.3), as

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} = \liminf_{T\to\infty} \frac{\sum_{i\neq 1} \Delta_i \mathbb{E}[N_i(T)]}{\log T} \geq \sum_{i\neq 1} \frac{\Delta_i}{d(\theta_i, \theta_1)}. \tag{2.4}$$

Lai and Robbins [1985] also showed that it is possible to construct an algorithm that is asymptotically optimal: that is, they showed a way to construct some statistical values that leads to an algorithm whose regret asymptotically matches the lower bound of inequality (2.4).

The above result was later extended in several directions. (i) More general class of reward distributions: Burnetas and Katehakis [1996] showed that it could be extended to multi-parameter distributions. Honda and Takemura [2010] showed that its bound could be applied to distributions with finite support. (ii) Even more general systems: Graves and Lai [1997] showed that the asymptotic lower bound can be extended to Markovian systems with a compact parameter space. Most papers on the bandit problem and its extensions implicitly or explicitly assume the strong consistency defined above. For example, algorithms that have logarithmic regret for any set of model parameters of interest (i.e., UCB1) are implicitly strongly consistent.

2.4 Adversarial Approach

Unlike the aforementioned approaches, the adversarial approach assumes nothing but the boundedness of the distribution of rewards. The bandit problem is formalized as a two-player game played by a forecaster and an adversary. The forecaster seeks a high reward, whereas the adversary tries to deceive the forecaster. Figure 2.4 formalizes the problem. The forecaster selects an arm (possibly using randomization), and at the same time, the adversary determines the reward of each arm in some bounded region to maximize the regret. Without loss of generality, we can assume the reward is in $[0, 1]$. The regret in this problem is defined as the difference between the cumulative reward of a single arm and the one of the algorithm.

$$\mathrm{Reg}(T) = \max_{i\in[K]} \left(\sum_{t=1}^{T} \widehat{X}_i(t) \right) - \sum_{t=1}^{T} \widehat{X}_{I(t)}(t).$$

The objective of the forecaster here is again to minimize the expected regret against the worst adversary who determines reward vectors based on the past selection of arms.

It is not difficult to show that any deterministic algorithm has linear regret: let the

Input: K (number of arms)
At each round $t = 1, ..., T$:

1. The forecaster selects an arm $I_t \in [K]$ (possibly using randomization).
2. The adversary at the same time selects the rewards of arms $\widehat{X}(t) = (\widehat{X}_1(t), \widehat{X}_2(t), \ldots, \widehat{X}_K(t))$.
3. The forecaster receives a reward $\widehat{X}_{I(t)}(t)$.

Goal: minimize the regret $\mathrm{Reg}(T) = \max_{i\in[K]} \left(\sum_{t=1}^{T} \widehat{X}_i(t)\right) - \sum_{t=1}^{T} \widehat{X}_{I(t)}(t)$.

Fig. 2.4. The adversarial bandit problem

algorithm be deterministic, and the adversary allocates the rewards as

$$\widehat{X}_i(t) = \begin{cases} 0 & (i = I(t)) \\ 1 & (\text{otherwise}). \end{cases} \tag{2.5}$$

As a result, the forecaster receives zero reward. In contrast, as the sum of the rewards allocated to the arms is $(K-1)T$, at least one of the arms has a cumulative regret larger than $(K-1)T/K$, which implies the regret is $\Omega(T)$.

The key in competing with an adversary is randomization: it is still possible to build an unbiased estimator of the cumulative reward of the arms. Let $p_i(t)$ be the probability that the arm i is selected at round t. Then,

$$\sum_{t=1}^{T} \frac{\widehat{X}_i(t)}{p_i(t)} \mathbf{1}\{I(t) = i\} \tag{2.6}$$

is an unbiased estimator of the cumulative reward of arm i. It is expected that, if the variance of this unbiased estimator is bounded as $o(T^2)$, we can make the regret sublinear: a class of randomized algorithms, called exponentially weighted forecasters, is known to have sublinear regret [Auer et al., 1995, 2002b]. Exp3 (Algorithm 1) is the most well-known version of the exponentially weighted forecaster. The following theorem states that the regret of Exp3 is upper bounded as $O(\sqrt{KT \log K})$.

Theorem 1. ([Auer et al., 2002b]) *The regret of Exp3 with* $\gamma = \min\left(1, \sqrt{\frac{K \log K}{(e-1)T}}\right)$ *is bounded as follows:*

$$2\sqrt{e-1}\sqrt{TK \log K}, \tag{2.7}$$

where $e \approx 2.73$ *is the base of the natural logarithm.*

It is also interesting to consider an algorithm that has works well in both stochastic and adversarial settings. Some studies [Bubeck and Slivkins, 2012, Seldin et al., 2014] have

Algorithm 1 Exp3 Algorithm [Auer et al., 2002b]

Input: # of arms K, $\gamma \in \mathbb{R}^+$
$t \leftarrow 1$, and $w_i(1) = 1$ for $i \in [K]$.
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, K$ **do**
    $p_i(t) \leftarrow (1-\gamma)\frac{w_i(t)}{\sum_{i=1}^{K} w_i(t)} + \frac{\gamma}{K}$
  end for
  Select $I(t)$ randomly in accordance with the probability $\{p_i(t)\}_{i\in[K]}$.
  Receive reward $\widehat{X}_{I(t)} \in [0,1]$.
  for $i = 1, 2, \ldots, K$ **do**
    $$\widehat{x}_i(t) = \begin{cases} \widehat{X}_i(t)/p_i(t) & (i = I(t)) \\ 0 & (\text{otherwise}) \end{cases}$$
    $w_i(t+1) \leftarrow w_i(t) \exp\left(\gamma \widehat{x}_i(t)/K\right)$
  end for
end for

proposed algorithms that simultaneously achieve good performance in both settings[*2].

2.5 Comparison of the Three Approaches

In this thesis, we design algorithms based on the stochastic approach, because it has the following advantages when it is applied to web systems:

- Algorithms based on the discounted cumulative reward are optimized for short-term rewards and explore too much when the discount factor is not properly set. Modern web systems involve a large number of people: the number of rounds T can be very large, and we can expect that the future reward will be as important as the current reward. In this sense, the undiscounted setting is more appropriate.
- The stochastic algorithms perform well with any parameter. In this sense, algorithms optimized for the stochastic approach are robust.
- The adversarial algorithms, such as Exp3, are also robust in regard to the reward distribution and balance exploration and exploitation in the game-theoretic sense. However, they are excessively conservative (i.e., they do more exploration than they need) when the reward is close to being stationary distributed.

A modern web server usually deals with a large number of users, and it makes sense to exploit the statistical properties of the rewards. The i.i.d. property of the rewards is the

[*2] More formally, such algorithms have $\tilde{O}(K)$ regret in the stochastic bandit and $\tilde{O}(\sqrt{KT})$ regret in the adversarial bandit, where $\tilde{O}$ hides a polylog factor as a function of K, T.

primary assumption placed on learning in many-user environments.

2.6 Proof of Asymptotic Regret Lower Bound

Here, we give a proof of the asymptotic regret lower bound in the stochastic approach (i.e., inequality (2.4)). For ease of analysis, the following theorem exclusively deals with the case of Bernoulli rewards. In this case, the parameter of each arm i corresponds to the expectation μ_i.

Theorem 2. (Asymptotic regret lower bound of a strongly consistent algorithm) *For any model parameters $\{\mu_i\}$ and any strongly consistent algorithm, the regret is lower bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \sum_{i \neq 1} \frac{(1 - o(1))\Delta_i}{d(\mu_i, \mu_1)} \log T,$$

where $d(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$ is the KL divergence between two Bernoulli distributions with parameters p and $q \in (0, 1)$.

To prove Theorem 2, we prove there is an asymptotic lower bound on the number of draws on each arm $i \neq 1$. The following proof follows essentially the same steps as Theorem 2 in Lai and Robbins [1985].

Lemma 3. *For any model parameters $\{\mu_i\}$ and any strongly consistent algorithm, the following inequality holds:*

$$\forall_{i \neq 1}\, \mathbb{E}[N_i(T)] \geq (1 - o(1)) \frac{\log T}{d(\mu_i, \mu_1)}. \tag{2.8}$$

Let $\widehat{\mu}_i$ be the empirical mean of the expected reward of arm i. Strong consistency requires an algorithm to make sure the possible risk that action $i \neq 1$ is optimal is smaller than $1/t$. The large deviation principle states that the probability that the arm with true parameter μ_i behaves like the arm of parameter $\widehat{\mu}_i$ is about $\exp\left(-N_i(T) d(\widehat{\mu}_i, \mu_i)\right)$. Therefore, we need to continue the exploration of arm i until $N_i(T) d(\mu_i, \mu_1) \sim \log t$ holds in order to reduce the risk that the expectation of arm i is larger than the one of arm 1 to $\exp\left(-\log t\right) = 1/t$.

Proof of Lemma 3. Let arm $i \neq 1$ be arbitrary suboptimal. Consider a modified bandit instance in which the expectation of arm i is different. Namely, the expectation of arm i is $\mu_i' > \mu_1$ such that

$$d(\mu_i, \mu_i') = d(\mu_i, \mu_1) + \epsilon. \tag{2.9}$$

From the monotonicity and continuity of the KL divergence, such a μ_i' uniquely exists for sufficiently small $\epsilon > 0$. The expectation of arm $j \neq i$ is the same as the one in the

original bandit instance. Note that, unlike the original bandit instance, the optimal arm in the modified bandit instance is not arm 1, but arm i.

Now, let $\widehat{X}_i^m \in \{0,1\}$ be the result of the m-th draw of arm i,

$$\widehat{\mathrm{KL}}(n) = \sum_{m=1}^{n} \log\left(\frac{\widehat{X}_i^m \mu_i + (1-\widehat{X}_i^m)(1-\mu_i)}{\widehat{X}_i^m \mu_i' + (1-\widehat{X}_i^m)(1-\mu_i')}\right),$$

and $\mathbb{P}'$, $\mathbb{E}'$ be the probability and expectation with respect to the modified bandit game. Let us define the events, Let us define the events

$$\begin{aligned}
\mathcal{D}_1 &= \left\{N_i(T) d(\mu_i, \mu_i') < (1-\epsilon)\log T, N_i(T) < \sqrt{T}\right\}, \\
\mathcal{D}_2 &= \left\{\widehat{\mathrm{KL}}(N_i(T)) \le \left(1-\frac{\epsilon}{2}\right)\log T\right\}, \\
\mathcal{D}_{12} &= \mathcal{D}_1 \cap \mathcal{D}_2, \\
\mathcal{D}_{1\setminus 2} &= \mathcal{D}_1 \cap \mathcal{D}_2^c.
\end{aligned}$$

First step ($\mathbb{P}\{\mathcal{D}_{12}\} = o(1)$): let $\{n_j\} \in \mathbb{N}^K$. We have,

$$\begin{aligned}
\mathbb{P}'(\mathcal{D}_{12} \cap \bigcap_{j\in[K]} \{N_j(T) = n_j\}) &= \int_{\mathcal{D}_{12}\cap\bigcap_{j\in[K]}\{N_j(T)=n_j\}} \exp\left(-\widehat{\mathrm{KL}}(N_i(T))\right) d\mathbb{P} \\
&\ge \mathbb{E}\left[\mathbf{1}\left\{\mathcal{D}_{12} \cap \bigcap_{j\in[K]} \{N_j(T) = n_j\}\right\} \exp\left(-\left(1-\frac{\epsilon}{2}\right)\log T\right)\right] \\
&= T^{-(1-\epsilon/2)} \mathbb{P}\left[\mathcal{D}_{12} \cap \bigcap_{j\in[K]} \{N_j(T) = n_j\}\right].
\end{aligned}$$

Summing over a disjoint union of events $\bigcap_{j\in[K]}\{N_j(T) = n_j\}$ for each $\{n_j\} \in \mathbb{N}^K$, we obtain

$$\mathbb{P}'(\mathcal{D}_{12}) \ge T^{-(1-\epsilon/2)} \mathbb{P}(\mathcal{D}_{12}).$$

Accordingly, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{D}_{12}) &\le T^{(1-\epsilon/2)} \mathbb{P}'(\mathcal{D}_{12}) \\
&\le T^{(1-\epsilon/2)} \mathbb{P}'\left(N_i(T) < \sqrt{T}\right) \\
&= T^{(1-\epsilon/2)} \mathbb{P}'\left(T - N_i(T) > T - \sqrt{T}\right) \\
&\le T^{(1-\epsilon/2)} \frac{\mathbb{E}'[T - N_i(T)]}{T - \sqrt{T}} \qquad \text{(by the Markov inequality).} \qquad (2.10)
\end{aligned}$$

Since this algorithm is strongly consistent, $\mathbb{E}'[T - N_i(T)] \to o(T^a)$ for any $a > 0$. Therefore, the RHS of the last line of (2.10) is $o(T^{a-\epsilon/2})$, which, by choosing a sufficiently small a, converges to zero as $T \to \infty$. In summary, $\mathbb{P}\{\mathcal{D}_{12}\} = o(1)$.

Second step ($\mathbb{P}\{\mathcal{D}_{1\backslash 2}\} = o(1)$): we have

$$\begin{aligned}
&\mathbb{P}\{\mathcal{D}_{1\backslash 2}\} \\
&= \mathbb{P}\left\{N_i(T)d(\mu_i,\mu_i') < (1-\epsilon)\log T, N_i(T) < \sqrt{T}, \widehat{\mathrm{KL}}(N_i(T)) > \left(1-\frac{\epsilon}{2}\right)\log T\right\} \\
&\leq \mathbb{P}\left\{\max_{n\in\mathbb{N}, d(\mu_i,\mu_i')<(1-\epsilon)\log T} \widehat{\mathrm{KL}}(n) > \left(1-\frac{\epsilon}{2}\right)\log T\right\}.
\end{aligned}$$

Note that

$$\max_{1\leq n\leq N} \widehat{\mathrm{KL}}(n) = \max_{1\leq n\leq N} \sum_{m=1}^{n} \log\left(\frac{\widehat{X}_i^m\mu_i + (1-\widehat{X}_i^m)(1-\mu_i)}{\widehat{X}_i^m\mu_i' + (1-\widehat{X}_i^m)(1-\mu_i')}\right),$$

is the maximum of the sum of positive-mean random variables, and thus it converges to its average (c.f., Lemma 10.5 in Bubeck, 2010). Namely,

$$\lim_{N\to\infty} \max_{1\leq n\leq N} \frac{\widehat{\mathrm{KL}}(n)}{N} = d(\mu_i,\mu_i') \quad \text{a.s.} \tag{2.11}$$

By using the fact that (2.11) holds almost surely and $1-\epsilon/2 > 1-\epsilon$, we have

$$\mathbb{P}\left(\max_{n\in\mathbb{N}, nd(\mu_i,\mu_i')<(1-\epsilon)\log T} \widehat{\mathrm{KL}}(n) > \left(1-\frac{\epsilon}{2}\right)\log T\right) = o(1).$$

In summary, we obtain $\mathbb{P}\left\{\mathcal{D}_{1\backslash 2}\right\} = o(1)$.

Last step: we have

$$\begin{aligned}
\mathcal{D}_1 &= \{N_i(T)d(\mu_i,\mu_i') < (1-\epsilon)\log T\} \cap \left\{N_i(T) < \sqrt{T}\right\} \\
&= \{N_i(T)(d(\mu_i,\mu_1)+\epsilon) < (1-\epsilon)\log T\} \cap \left\{N_i(T) < \sqrt{T}\right\} \qquad \text{(By (2.9))} \\
&\supseteq \left\{N_i(T)(d(\mu_i,\mu_1)+\epsilon) + \frac{(1-\epsilon)\log T}{\sqrt{T}} N_i(T) < (1-\epsilon)\log T\right\},
\end{aligned}$$

where we have used the fact that $\{A < C\} \cap \{B < C\} \supseteq \{A + B < C\}$ for $A, B > 0$ in the last line. Note that, by using the result of the previous steps, $\mathbb{P}\{\mathcal{D}_1\} = \mathbb{P}\{\mathcal{D}_{12}\} + \mathbb{P}\{\mathcal{D}_{1\backslash 2}\} = o(1)$. By using the complement of this fact, we have

$$\mathbb{P}\left\{N_i(T)(d(\mu_i,\mu_1)+\epsilon) + \frac{(1-\epsilon)\log T}{\sqrt{T}} N_i(T) \geq (1-\epsilon)\log T\right\} \geq \mathbb{P}\{\mathcal{D}_1^c\} = 1 - o(1).$$

The Markov inequality yields

$$\mathbb{E}\left\{N_i(T)(d(\mu_i,\mu_1)+\epsilon) + \frac{(1-\epsilon)\log T}{\sqrt{T}} N_i(T)\right\} \geq (1-\epsilon)(1-o(1))\log T. \tag{2.12}$$

Because $\mathbb{E}[N_i(T)]$ is a subpolynomial function of T due to consistency, the second term in the LHS of (2.12) is $o(1)$ and thus negligible. Lemma 27 follows from the fact that (6.11) holds for sufficiently small ϵ. □

Proof of Theorem 2. The proof follows directly from Lemma 3:

$$\mathbb{E}[\mathrm{Reg}(T)] = \sum_{i \in [K]} \Delta_i \mathbb{E}[N_i(t)] \geq (1 - o(1)) \frac{\Delta_i \log T}{d(\mu_i, \mu_1)}.$$

□

Chapter 3

Algorithms for Solving Multi-armed Bandit Problem

Chapter 2 explained the three approaches to the bandit problem. The rest of this thesis concerns the stochastic approach, which exploits the i.i.d. property of the rewards. The asymptotic regret lower bound, which defines the optimality of an algorithm under the strong consistency assumption, was derived in Section 2.3.2. In this chapter, we discuss the well-known stochastic bandit algorithms and the ideas underlying them. Table 3.1 compares these algorithms, i.e., ϵ-greedy, Upper Confidence Bound (UCB), Thompson sampling (TS), and Deterministic Minimum Empirical Divergence (DMED). ϵ-greedy is a popular heuristic algorithm that is used in the field of reinforcement learning (Section 3.1). However, ϵ-greedy and its versions are not asymptotically optimal in terms of regret: whereas it conducts uniform exploration over all suboptimal arms, an asymptotically optimal algorithm needs to control the amount of exploration adaptively for each arm. The other three algorithms can adaptively control the number of draws and obtain an asymptotically optimal regret (Section 3.2). Each algorithm has its own way of determining the next arm to draw. UCB is a class of algorithms that explicitly follows the

Table. 3.1. Multi-armed bandit algorithms. An algorithm is defined to be asymptotically optimal if it has a regret bound that asymptotically matches the lower bound of Section 2.3.2.

Algorithm	ϵ-greedy	UCB	TS	DMED
asymptotically optimal	no	yes	yes	yes
strategy	uniform exploration	optimistic within confidence interval	posterior sampling	likelihood-based exploration

idea of the optimism under uncertainty (Section 3.3). TS [Thompson, 1933] is an old heuristic for sequential decision-making on the sampling among distributions, which was first proposed in the 1930s (Section 3.4). TS takes a prior and is inherently Bayesian. DMED [Honda and Takemura, 2010] is a recent algorithm that is based on the likelihood of each arm being the optimal one (Section 3.5). We will analyze the regret bounds of these algorithms. For ease of analysis, we will restrict ourselves to the Bernoulli bandit problem, where rewards are in $\{0, 1\}$. Overall, most of the results in this section can be extended to a one-parameter canonical exponential family of reward distributions, such as Gaussian distributions with a known variance. Section 3.6 discusses the performance of the algorithms, while Section 3.7 discusses other topics, such as extendability. The notation in this chapter is summarized in Table 3.2.

Table. 3.2. Notation used in Chapter 3.

$\mathbf{1}\{A\}$	:=	1 if A is true and 0 otherwise.
K	:=	Number of the arms.
$[K]$	:=	$\{1, 2, \ldots, K\}$.
T	:=	Number of the rounds.
$I(t)$	:=	The arm that is selected in round t.
$\widehat{X}_i(t)$	:=	Reward of arm i at round t.
μ_i	:=	Mean reward of arm i.
Δ_i	:=	$\mu_1 - \mu_i$.
$\widehat{\mu}_i(t)$	:=	Empirical mean reward of arm i at round t.
$\widetilde{\mu}_i(t)$	:=	Posterior sample of μ_i at round t in Thompson sampling.
$N_i(t)$	:=	Number of rounds in which arm i is selected before round t: that is, $\sum_{t'=1}^{t-1} \mathbf{1}\{I(t') = i\}$.
$d(p, q)$	:=	$p \log(p/q) + (1-p) \log((1-p)/(1-q))$.
$F^{\mathrm{beta}}_{\alpha,\beta}(y)$	:=	Cumulative distribution function of the beta distribution with integer parameters α and β.
$F^{\mathrm{B}}_{n,p}(\cdot)$	:=	Cumulative distribution function of the binomial distribution with parameters n, p.

Large deviation inequalities: the stochastic bandit problem involves an estimation of the expectation of each arm. Therefore, we need to know the probability of the deviation of the empirical means of the arms from the true mean. In this thesis, we use the following large deviation inequalities that bound the tail probability of the sum of random variables. Hoeffding's inequality holds for random variables with finite support, and the Chernoff bound holds for binary variables.

Fact 4. (Hoeffding's inequality for random variables with finite support)

Algorithm 2 ϵ-greedy and ϵ_t-greedy

Input: # of arms K, $\epsilon \in (0,1)$ (ϵ-greedy), $c \in \mathbb{R}^+$ (ϵ_t-greedy)

for $t = 1, 2, \ldots, T$ **do**

Let $\widehat{i}^*$ be the arm with the largest empirical mean reward (ties broken arbitrarily).

$\epsilon \leftarrow \min(1, c/t)$ (ϵ_t-greedy)

With probability $1-\epsilon$, select $\widehat{i}^*$ as $I(t)$. With probability ϵ, select $I(t)$ uniformly at random.

end for

Let $\widehat{X}_1, \ldots, \widehat{X}_n$ be i.i.d. random variables on $[0,1]$, and let $\widehat{X} = \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i$ and $\mu = \mathbb{E}[\widehat{X}_i]$. Then, for any $\epsilon \in \mathbb{R}^+$,

$$\mathbb{P}(\widehat{X} \geq \mu + \epsilon) \leq \exp\left(-2\epsilon^2 n\right).$$

and for any $\epsilon \in \mathbb{R}^+$,

$$\mathbb{P}(\widehat{X} \leq \mu - \epsilon) \leq \exp\left(-2\epsilon^2 n\right).$$

The union bound of the two inequalities yields

$$\mathbb{P}(|\widehat{X} - \mu| \geq \epsilon) \leq 2\exp\left(-2\epsilon^2 n\right).$$

Fact 5. (Chernoff bound for binary random variables)

Let $\widehat{X}_1, \ldots, \widehat{X}_n$ be i.i.d. binary random variables. Let $\widehat{X} = \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i$ and $\mu = \mathbb{E}[\widehat{X}_i]$. Then, for any $\epsilon \in \mathbb{R}^+$,

$$\mathbb{P}(\widehat{X} \geq \mu + \epsilon) \leq \exp\left(-d(\mu+\epsilon, \mu)n\right).$$

and for any $\epsilon \in \mathbb{R}^+$,

$$\mathbb{P}(\widehat{X} \leq \mu - \epsilon) \leq \exp\left(-d(\mu-\epsilon, \mu)n\right).$$

3.1 ϵ-greedy

ϵ-greedy (Algorithm 2) is a well-known algorithm for solving reinforcement learning problems [Sutton and Barto, 1998]. The idea behind it is quite simple: it exploits on the basis of current knowledge with probability $1-\epsilon$ and explores with probability ϵ. One can expect the exploration and exploitation trade-off can be balanced by optimizing the probability ϵ. Unfortunately, this algorithm is not optimal for the multi-armed bandit problem since an $\epsilon \in (0,1)$ fraction of rounds generates regret, and thus, its regret is linear to the number of the rounds. A version of this algorithm, called ϵ_t-greedy [Auer

et al., 2002a], is also well-known in the field of machine learning. ϵ_t-greedy decreases the probability of exploration at an $O(1/t)$ rate, which makes sense because the optimal amount of exploration in the multi-armed bandit problem is $O(\log T)$ and $\sum_{t=1}^{T}(1/t) = O(\log T)$. The following theorem implies that, by choosing a sufficiently large c, ϵ_t-greedy has a logarithmic regret bound.

Theorem 6. (Theorem 3 in Auer et al. [2002a]) *Let $\Delta_2 := \mu_1 - \mu_2 > 0$. For an ϵ_t-greedy algorithm with $c > \max(5K, (2K)/(\Delta_2)^2)$, for all $t > c\Delta_2$, the probability that the algorithm selects a suboptimal arm is $\Theta(1/t)$.*

Sketch of the proof of Theorem 6. Uniform exploration ($\epsilon_t = c/t$) is sufficient if we select a sufficiently large parameter c. The appropriate value of c depends on Δ_2. This is because the expectation of arm 2 is the closest to the one of arm 1 (i.e., the optimal arm): it is arm 2 that requires the largest number of samples to be distinguished from arm 1.

Let arm $i \neq 1$ be arbitrary suboptimal. We have

$$\mathbb{P}[I(t) = i] \leq \mathbb{P}[\widehat{\mu}_i(t) \geq \mu_i + \Delta_i/2] + \mathbb{P}[\widehat{\mu}_1(t) \leq \mu_1 - \Delta_i/2]. \tag{3.1}$$

In what follows, we bound the first term of the RHS. The second term of the RHS can be bounded by using the same argument. The first term is decomposed as

$$\begin{aligned}
&\mathbb{P}[\widehat{\mu}_i(t) \geq \mu_i + \Delta_i/2] \\
&\quad \leq \underbrace{\sum_{n=0}^{\lfloor x_0 \rfloor} \mathbb{P}[\widehat{\mu}_i(t) \geq \mu_i + \Delta_i/2, N_i(t) = n]}_{(X)} + \underbrace{\sum_{n=\lceil x_0 \rceil}^{\infty} \mathbb{P}[\widehat{\mu}_i(t) \geq \mu_i + \Delta_i/2, N_i(t) = n]}_{(Y)},
\end{aligned}$$

where $x_0 = (c/2K)\sum_{t=1}^{T}$. From Bernstein's inequality, one can prove that uniform exploration for each $\mathbb{P}[N_i(t) = n] \leq \exp^{-x_0/5}$, which, with the union bound over $n = 1, \ldots, x_0$, bounds term (X). Assuming that the arm is drawn $N_i(t) = n$ times, from Hoeffding's inequality, it follows that

$$\mathbb{P}[\widehat{\mu}_{i,n} - \mu_i \geq \Delta_i/2] \leq \mathrm{e}^{-\Delta_i^2 n/2},$$

which, by summing over $n \geq \lceil x_0 \rceil$, bounds term (Y). In summary, term (X) is $O(x_0 \mathrm{e}^{-x_0/5})$, and term (Y) is $O(\mathrm{e}^{-\Delta_i^2 x_0/2})$. The first term on the RHS of (3.1) is upper bounded by the sum of terms (X) and (Y). Theorem follows by lower-bounding x_0 as

$$x_0 \geq \frac{c}{K} \log \frac{T}{c}.$$

□

Although the logarithmic regret bound directly follows from Theorem 6, the performance of the algorithm depends on the constant c. The optimal value of c that minimizes

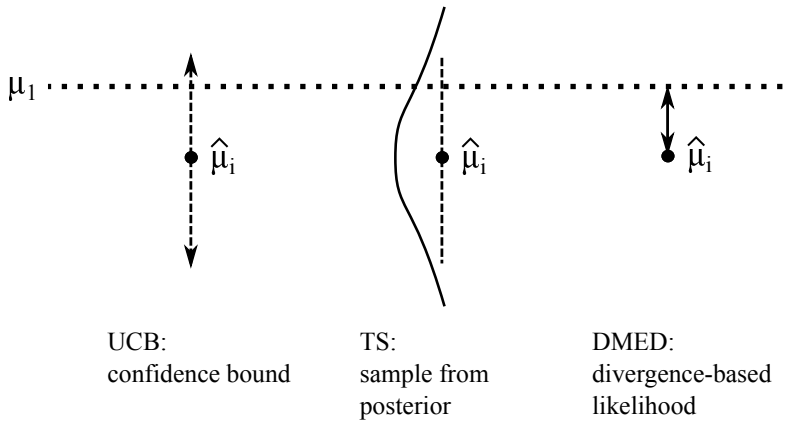


Fig. 3.1. Illustration of the three asymptotically optimal algorithms. UCB uses the top $(1/t)$-quantile of the confidence bound, which in practice is approximated by a large deviation bound. TS samples from the posterior distribution. DMED is based on the empirical divergence, the negative exponential of which can be considered as a likelihood that the expectation of arm i is larger than the empirical best one. DMED continues exploring until the likelihood becomes $\sim 1/t$.

the regret depends on the model parameters, which are hard to optimize beforehand. In general, ϵ_t-greedy is not asymptotically optimal, because it explores all arms uniformly: the sufficient amount of exploration differs among arms, and thus, an optimal algorithm needs to control the amount of exploration for each arm. This aspect is not considered in greedy algorithms.

Despite the lack of a strong theoretical guarantee, ϵ-greedy and ϵ_t-greedy are widely used in practice. Examples of applications of ϵ-greedy include the following. Li et al. [2010] tested ϵ-greedy for news article recommendation on the front page of Yahoo! Banditron [Valizadegan et al., 2011], a bandit-based algorithm for multi-class prediction with partial feedback, balances exploration and exploitation by using the ϵ-greedy strategy. Motivated by e-commerce applications, Chakrabarti et al. [2008] proposed a version of the bandit problem in which each arm has a lifetime after which it is no longer available. They proposed adaptive-greedy heuristics, a variant of ϵ-greedy that adaptively tunes ϵ based on the past rewards.

3.2 Asymptotically Optimal Algorithms: Controlling the Number of Draws

In Section 3.1, we introduced ϵ-greedy and its variant ϵ_t-greedy. Although ϵ_t-greedy has a logarithmic regret, the leading constant in front of the logarithmic factor is hard

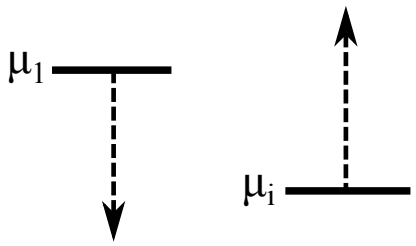


Fig. 3.2. The draw of suboptimal arm i implies either (i) underestimation of the optimal arm or (ii) overestimation of arm i.

to optimize because ϵ_t-greedy conducts a uniform exploration over all arms. In contrast, UCB, TS, and DMED that we explain later are asymptotically optimal in the sense of the regret. In this section, we explain the underlying idea of these algorithms. The asymptotic regret lower bound (c.f., Section 2.3.2) is the sum over $K-1$ suboptimal arms. To confirm that arm i is not as good as the optimal one, a strongly consistent bandit algorithm needs to draw each arm

$$\frac{\log T}{d(\mu_i, \mu_1)}$$

times. Of course, an algorithm is not informed of the model parameters $\{\mu_i\}_{i\in[K]}$, and thus it needs to adapt to the distributions. In the following, we give an intuitive argument as to why an asymptotically optimal algorithm can control the number of draws. An algorithm with sublinear regret selects the optimal arm in most of the rounds: suppose that $N_i(t)$ for each suboptimal arm $i \neq 1$ is $o(t)$. This implies $N_1(t) = t - \sum_{i\neq 1} N_i(t) = t - o(t)$. Given $O(t)$ samples, the deviation $|\widehat{\mu}_1 - \mu_1|$ is $O(1/\sqrt{t})$ from the law of large numbers, which is very small compared with $|\widehat{\mu}_i - \mu_i|$ of arm $i \neq 1$. For this reason, we can identify $\widehat{\mu}_1$ and μ_1 when we examine the leading logarithmic term. To satisfy consistency, an algorithm needs to check that the true expectation of arm i is less than that of arm 1 with a confidence level of $\sim 1/t$. Assume that arm i has the expectation $\mu_i' = \mu_1$, but its empirical expectation $\widehat{\mu}_i \sim \mu_i$. The large deviation principle states that $N_i(t) \sim \frac{\log t}{d(\mu_i,\mu_1)}$ implies $\mathbb{P}\{\widehat{\mu}_i(t) < \mu_i\} \leq \exp\left(-N_i(t)d(\mu_i, \mu_1)\right) \sim 1/t$ if arm i's expectation is μ_1.

The idea behind the three algorithms is illustrated in Figure 3.1: to approximate a $1/t$ confidence level, UCB, TS, and DMED respectively use an upper confidence bound, posterior samples, or a likelihood function. These algorithms are described in the later sections.

3.2.1 Regret analysis: general idea

Given an algorithm, our next concern is how to analyze its regret. The technique depends on the algorithm, yet we can discern that they share a common structure: the draw of arm $i \neq 1$ implies either an underestimation of the optimal arm or an overestimation

of arm i. Namely,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbf{1}\{I(t) = i\} \\
&\subset \sum_{t=1}^{T} \mathbf{1}\{\text{Underestimation of the optimal arm}\} + \sum_{t=1}^{T} \mathbf{1}\{\text{Overestimation of arm } i\} \\
&\subset \sum_{t=1}^{T} \mathbf{1}\{\text{Underestimation of the optimal arm}\} \\
&\quad + \sum_{t=1}^{T} \mathbf{1}\{\text{Overestimation of arm } i \text{ after it is sufficiently sampled.}\} \\
&\quad + (\text{sufficient number of samples: } \frac{\log T}{d(\mu_i, \mu_1)} + o(\log T)),
\end{aligned}$$

which is illustrated in Figure 3.2. The first term is the underestimation of the optimal arm. Since the optimal arm is often sampled, the overestimation of this term is unlikely to happen and is expected to be $o(\log T)$. Interestingly, bounding this term is often the most difficult part of the proof. The second term is the overestimation of arm i. Given sufficient samples, the use of large deviation inequalities leads to the bound of this term. Therefore, the overestimation of arm i after a sufficient number of samples is expected to be $o(\log T)$. Below, we discuss the individual algorithms.

3.3 Upper Confidence Bound (UCB)

UCB is a class of algorithms that is based on the upper confidence bound on the expected reward of each arm. UCB is a realization of the "optimism under uncertainty" principle: the uncertainty of each arm is optimistically evaluated in the form of the upper confidence bound, which turns out to be effective on bandit problems. There are several versions of UCB: probably the most famous one is UCB1 [Auer et al., 2002a] (Algorithm 3). However, the use of an upper confidence bound in bandit problems goes back at least to the 1980s [Lai and Robbins, 1985].

Theorem 7. (Regret bound of UCB1) *The regret of UCB1 with $\alpha > 1$ is bounded as*

$$\mathrm{Reg}(T) \leq \sum_{i \neq 1} \frac{4\alpha \log T}{\Delta_i} + O(1),$$

where $O(1)$ is a constant as a function of T.

UCB1 has a logarithmic regret, but its constant factor on the logarithmic term is not optimal because its confidence bound is based on Hoeffding's inequality. Although Hoeffding's inequality holds for any reward distribution with finite support, it is not very tight on a specific family of distributions, such as Bernoulli.

Algorithm 3 UCB1

Input: # of arms K, $\alpha \in \mathbb{R}^+$.

Select each arm once and receive the rewards. $t \leftarrow K+1$.

for $t = K+1, 2, \ldots, T$ **do**

Calculate the UCB1 index of each arm as

$$c_i(t) = \widehat{\mu}_i(t) + \sqrt{\frac{\alpha \log t}{N_i(t)}}$$

Select the arm of the largest UCB1 index as $I(t)$ (ties broken arbitrarily).

end for

Relation to the discussion in Section 3.2.1: Section 3.2.1 indicates that deriving the bound on the underestimation of the optimal arm is usually the hardest part in the regret analysis. The analysis of this term in UCB1 is easier because Hoeffding' s inequality approximates the KL divergence, and thus, it does not require the inverse of the KL divergence to be calculated.

Relation between the $O(\log T/\Delta_i)$ bound and the optimal bound: the leading constant $4\alpha \log T \sum_{i \neq 1}(1/\Delta_i)$ can be considered to be an approximation of the asymptotically optimal bound in the following sense. Pinsker's inequality (inequality (35)) states that

$$d(\mu_i, \mu_1) \geq 2\Delta_i^2, \tag{3.2}$$

and thus, the regret of UCB1 is larger than the optimal bound:

$$\sum_{i \neq 1} \frac{\Delta_i \log T}{d(\mu_i, \mu_1)} \leq \sum_{i \neq 1} \frac{\log T}{2\Delta_i} < 4\alpha \sum_{i \neq 1} \frac{\log T}{\Delta_i}. \tag{3.3}$$

On the other hand, $d(\mu_i, \mu_1)$ is upper-bounded as follows:

$$\begin{aligned} d(\mu_i, \mu_1) &= \int_{\mu_i}^{\mu_1} \frac{\partial d(\mu_i, x)}{\partial x} dx \\ &= \int_{\mu_i}^{\mu_1} \frac{x - \mu_i}{(1-x)x} dx \\ &\leq \frac{\Delta_i^2}{(1-\mu_1)\mu_i}. \end{aligned}$$

Therefore, when $\min(\mu_i, 1-\mu_1)$ is not very small, $d(\mu_i, \mu_1) \sim \Theta(\Delta^2)$ is a good approximation. Unlike the regret bound of ϵ_t-greedy, which cannot adapt to model parameters $\{\mu_i\}$, the regret bound of UCB1 reflects the reward gap $\{\Delta_i\}$ and is more reasonable.

Because of its simplicity and reasonable regret bound, UCB1 has been widely used in the machine learning community. There are hundreds of applications and extensions, but we will not discuss them in any detail.

3.3.1 Analysis of UCB1

In this subsection, we give a proof of the UCB1 regret bound.

Proof of Theorem 7. To prove the theorem, it suffices to show that for any $i \neq 1$

$$\mathbb{E}[N_i(T+1)] \leq \frac{4\alpha \log T}{\Delta_i^2} + O(1), \tag{3.4}$$

since $\mathrm{Reg}(T) = \sum_{i \neq 1} \Delta_i N_i(T+1)$.

To simplify the discussion, we will ignore the initial exploration (one sample from each arm) since the regret in this duration is at most K. Let

$$\begin{aligned}
\mathcal{A}(t) &:= \left\{ \widehat{\mu}_1(t) + \sqrt{\frac{\alpha \log t}{N_1(t)}} < \mu_1 \right\} \\
\mathcal{B}(t) &:= \left\{ \widehat{\mu}_i(t) > \mu_i + \sqrt{\frac{\alpha \log t}{N_i(t)}} \right\} \\
\mathcal{C}(t) &:= \left\{ N_i(t) \leq \frac{4\alpha \log t}{\Delta_i^2} \right\}.
\end{aligned}$$

In the following, we often denote $\{\mathcal{A}, \mathcal{B}\}$ instead of $\{\mathcal{A} \cap \mathcal{B}\}$ for two events $\mathcal{A}$ and $\mathcal{B}$. $\mathcal{A}(t)$ means underestimation of arm 1 (optimal arm), $\mathcal{B}(t)$ means overestimation of arm i, and $\mathcal{C}(t)$ means that arm i is not sufficiently sampled.

Note that

$$\mathbf{1}\{I(t) = i\} \leq \mathbf{1}\{\mathcal{A}(t)\} + \mathbf{1}\{\mathcal{B}(t)\} + \mathbf{1}\{\mathcal{C}(t)\} \tag{3.5}$$

since $\{\mathcal{A}(t)^c \cap \mathcal{B}(t)^c \cap \mathcal{C}(t)^c\}$ implies

$$\widehat{\mu}_1(t) + \sqrt{\frac{\alpha \log t}{N_1(t)}} \geq \mu_1 = \mu_i + \Delta_i > \mu_i + 2\sqrt{\frac{\alpha \log t}{N_i(t)}} \geq \widehat{\mu}_i + \sqrt{\frac{\alpha \log t}{N_i(t)}},$$

and thus the arm i is not selected. Next, we bound each event. Let $\widehat{\mu}_{1,n}$ be the empirical mean of n samples on arm 1. We have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}\{\mathcal{A}(t)\} &\leq \sum_{t=1}^{T} \sum_{n=1}^{t} \mathbb{P}\left\{ \widehat{\mu}_1(t) + \sqrt{\frac{\alpha \log t}{N_1(t)}} < \mu_1, N_1(t) = n \right\} \\
&= \sum_{t=1}^{T} \sum_{n=1}^{t} \mathbb{P}\left\{ \widehat{\mu}_{1,n} + \sqrt{\frac{\alpha \log t}{n}} < \mu_1 \right\} \\
&\leq \sum_{t=1}^{\infty} \sum_{n=1}^{t} \frac{1}{t^{2\alpha}} \qquad \text{(by Hoeffding's inequality)} \\
&\leq \sum_{t=1}^{\infty} \frac{1}{t^{2\alpha-1}} \leq O(1).
\end{aligned}$$

where we have used $2\alpha - 1 > 1$ and $\sum_{t\in\mathbb{N}}(1/t^z) = O(1)$ for any $z > 1$. The same argument yields

$$\sum_{t=1}^{T} \mathbb{P}\{\mathcal{B}(t)\} = O(1).$$

The term of $\mathcal{C}(t)$ is bounded as

$$\begin{aligned}\sum_{t=1}^{T} \mathbf{1}\{\mathcal{C}(t)\} &\leq \frac{4\alpha \log T}{\Delta_i^2} + \mathbf{1}\left\{\mathcal{C}(t), N_i(t) > \frac{4\alpha \log T}{\Delta_i^2}\right\} \\ &= \frac{4\alpha \log T}{\Delta_i^2}.\end{aligned}$$

Taking the expectation of (3.5) and using the three bounds above yield (3.4). □

Note that a more refined analysis with Hoeffding's maximal inequality and the peeling trick slightly improves a bound (c.f., Theorem 2.2 in Bubeck [2010]).
Nevertheless, the leading logarithmic constant of UCB1 cannot be optimal from the fact that it uses Hoeffding's inequality to approximate the confidence bound.

3.3.2 KL-UCB: optimizing the leading coefficient

As we discussed in Section 3.2, UCB needs to set the confidence level to $1/t$ to be strongly consistent. The confidence level determined by the large deviation principle, which describes the tail probability of the distributions. UCB1 defines the confidence bound by using Hoeffding's inequality. Although the inequality holds for any finite support distribution, it is loose when we consider a family of specific distributions, such as Bernoulli distributions.

A more refined algorithm, called KL-UCB (Algorithm 4), approximates the $1/t$ confidence bound by using KL-divergence-based concentration inequalities[*1]. Note that the KL-UCB index can be efficiently calculated by using Newton or bisection iterations. Although KL-UCB was proposed in Garivier and Cappé [2011], essentially the same idea on the confidence bound was proposed in Lai [1987].

Section 3.3.1 showed that the underestimation of the optimal arm is $o(\log T)$. This still holds for KL-UCB. Unlike UCB1, the analysis of KL-UCB is rather involved. KL-UCB uses a KL-divergence-based confidence bound that is optimal with respect to its exponential factor. As a result, we need a more involved analysis of the self-normalized inequality to bound this factor. For more details, see Kaufmann [2014].

Theorem 8. (Theorem 1 in Cappé et al. [2013]) *For KL-UCB with $c = 3$ and for any suboptimal arm $i \neq 1$,*

$$\mathbb{E}[N_i(T+1)] \leq \frac{\log T}{d(\mu_i, \mu_1)} + o(\log T),$$

[*1] In the case of the Bernoulli distribution, this corresponds to the Chernoff bound (Fact 5).

Algorithm 4 KL-UCB

Input: # of arms K, $c \in \mathbb{R}^+$.

Select each arm once and receive the rewards. $t \leftarrow K+1$.

for $t = K+1, 2, \ldots, T$ **do**

Calculate the KL-UCB index of each arm as

$$c_i(t) = \left\{ \max_{q \in [0,1]} : N_i(T) d(\widehat{\mu}_i(t), q) \leq \log t + c \log(\log t) \right\}.$$

Select the arm of the largest KL-UCB index as $I(t)$ (ties broken arbitrarily).

end for

from which an asymptotically optimal bound on the regret directly follows.

3.4 Thompson Sampling (TS)

Thompson sampling is one of the oldest heuristics in the field of sequential decision-making. Motivated by clinical trials, Thompson [1933] originally studied the probability that one distribution is superior to another in the sense of its expectation. Although Thompson [1933] was originally interested in the computational aspect of this probability, today we can obtain the numerical value of such a probability relatively easily for many distributions, and here, we are more interested in the general idea of selecting arms in the bandit problem. Today, we use the term "Thompson sampling" to refer to a class of algorithms that behave greedily based on posterior samples. TS is inherently Bayesian: given a prior distribution over the parameters $P(\theta)$ and observations $\mathcal{D}_1, \ldots \mathcal{D}_t$ of the rewards of the selected arm, it computes the posterior distribution by using Bayes's rule $P(\theta|\mathcal{D}_1, \ldots, \mathcal{D}_t) \propto P(\mathcal{D}_1, \ldots, \mathcal{D}_t|\theta)P(\theta)$. A closed-form expression for the posterior distribution can be computed in the case of the conjugate prior. In the multi-armed bandit problem, the model parameter of each arm is disjoint. Algorithm 5 is TS with Bernoulli rewards. At the beginning, TS initializes the distribution of each arm with a uniform prior Beta$(1, 1)$. At each round, TS selects an arm and receives a reward. The posterior of each arm i is Beta(A_i, B_i), where $A_i = A_i(t)$ and $B_i = B_i(t)$ are one plus the number of rewards 1 and 0, respectively. For selecting an arm, TS adopts the posterior sampling method: namely, it selects each arm in accordance with the probability that it maximizes the expected reward; that is,

$$\int \mathbf{1}\{\mu_i = \max_{j \in [K]} \mu_j\} P(\mu|\mathcal{D}_1, \ldots, \mathcal{D}_t) d\mu.$$

An explicit computation of this probability is not required. In implementing TS, just a single sample from the posterior distribution is enough. TS for the Bernoulli bandit samples from each arm's posterior $\widetilde{\mu}_i(t) \sim$ Beta$(A_i(t), B_i(t))$ and chooses arm $\arg\max_i \widetilde{\mu}_i(t)$. Sam-

Algorithm 5 Thompson sampling (TS) for binary rewards.

Input: # of arms K
for $i = 1, 2, \ldots, K$ **do**
  $A_i, B_i \leftarrow 1, 1$
end for
$t \leftarrow 1$.
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, K$ **do**
    $\widetilde{\mu}_i(t) \sim \mathrm{Beta}(A_i, B_i)$
  end for
  $I(t) = \arg\max_i \widetilde{\mu}_i(t)$ (ties broken arbitrarily).
  if $\widehat{X}_{I(t)}(t) = 1$ **then**
    $A_{I(t)} \leftarrow A_{I(t)} + 1$
  else
    $B_{I(t)} \leftarrow B_{I(t)} + 1$
  end if
end for

pling from many posterior distributions such as the Bernoulli and Normal distributions is easy, and thus, Thompson sampling is computationally efficient.

Thompson sampling had long been forgotten until recently. Scott [Scott, 2010, 2015] found that TS is effective at optimizing web systems, and Chapelle and Li [2011] verified its effectiveness at online advertisement selection. Subsequently, people have analyzed its performance from a theoretical viewpoint. In the case of Bernoulli rewards, a TS with a uniform prior has been shown to be asymptotically optimal in the sense of stochastic regret: [Ortega and Braun, 2010] showed the regret is $o(t)$, and Agrawal and Goyal [2012] showed the regret is $O(\log T)$ with the introduction of techniques for Bernoulli rewards. Kaufmann et al. [2012], Agrawal and Goyal [2013a] showed its asymptotic optimality.

Theorem 9. (Theorem 1 in Agrawal and Goyal [2013a])
The regret of TS is upper bounded as

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \neq 1} \frac{\Delta_i \log T}{d(\mu_i, \mu_1)} + o(\log T).$$

It is interesting that this Bayesian algorithm is optimal in the sense of the frequentist interpretation. Later, Korda et al. [2013] showed that choosing the Jeffrey's prior leads to an asymptotically optimal regret for the one-parameter exponential family of distributions, which is a generalization of the earlier result on Bernoulli bandits. However, Jeffrey's prior is not always good if we consider multi-parameter distributions. In some multi-parameter

distributions, it is known that a naive choice of the prior does not suffice to have strong consistency: Honda and Takemura [2014] showed that in TS for Normal distributions with unknown mean and variance, using Jeffrey's prior can result in a linear regret. A recent trend in machine learning is to generalize the idea of TS; such studies include Agrawal and Goyal [2013b], Gopalan et al. [2014], Kocák et al. [2014].

3.4.1 Idea behind TS

As we discussed in Section 3.2, strong consistency requires that $N_i(t) \sim \frac{\log t}{d(\mu_i, \mu_1)}$. In other words, a suboptimal arm needs to be drawn until it is suboptimal with a confidence level of $1/t$. Here, we give an intuitive explanation on how TS controls the probability of selecting suboptimal arms. A more formal analysis is provided in Section 3.4.2.

TS selects each arm in accordance with the probability that it maximizes the expected reward. Let $p_i(t)$ be the probability that arm i is selected at round t. When t is large, $p_i(t)$ should be the approximate empirical probability that the arm was selected up to that round, which is expected to be close to $\mathbb{P}[\widetilde{\mu}_i(t) \geq \mu_1]$. That is,

$$N_i(t) \sim p_i(t)t. \tag{3.6}$$

Assuming that $N_i(t) \sim \log t = \tilde{\Theta}(1)$[*2], $p_i(t) \sim 1/t$. Let $F^{\mathrm{beta}}_{\alpha,\beta}(y)$ be the cdf of a beta distribution with integer parameters α and β. Furthermore, let $F^{\mathrm{B}}_{n,p}(\cdot)$ be the cdf of a binomial distribution with parameters n, p. Since TS selects an arm based on the posterior sample, $p_i(t)$ corresponds to the tail probability $1 - F^{\mathrm{beta}}_{n\widehat{\mu}_{1,n}+1, n(1-\widehat{\mu}_{1,n})+1}(\mu_1)$, where we define $n = N_i(t)$ to simplify the notation. From the Beta-Binomial equality, (Fact 34) turns out to be

$$1 - F^{\mathrm{beta}}_{n\widehat{\mu}_{1,n}+1, n(1-\widehat{\mu}_{1,n})+1}(\mu_1) = F^{\mathrm{B}}_{n+1,\mu_1}(n\widehat{\mu}_{1,n}). \tag{3.7}$$

Assuming that $\widehat{\mu}_{1,n} \sim \mu_i$, the RHS of (3.7) is $F^{\mathrm{B}}_{n+1,\mu_1}(n\mu_i)$. This quantity is the probability that the Bernoulli(μ_1) behaves like Bernoulli(μ_i), which is the confidence level. In summary, the confidence level of arm i being suboptimal is $\sim 1/t$.

3.4.2 Analysis of TS

In this section, we provide a sketch of the proof of the regret bound of TS. Although we focus on the analysis of Bernouli rewards, the technique here applies to many reward settings, such as the one-parameter exponential family [Korda et al., 2013] or the Normal distribution [Honda and Takemura, 2014]. The analysis is mainly based on the techniques in Agrawal and Goyal [2013a] and Honda and Takemura [2014].

[*2] $\tilde{\Theta}$ ignores a polylog factor.

Sketch of the proof of Theorem 9. We prove that, for an arbitrary arm $i \neq 1$ and sufficiently small $\epsilon > 0$,

$$\mathbb{E}[N_i(T+1)] \leq \frac{\log T}{(1-\epsilon)d(\mu_i, \mu_1)} + O(1), \tag{3.8}$$

from which Theorem 9 easily follows.

Let $\widehat{\mu}_i(t)$ be the empirical mean reward of arm i at round t and $\widetilde{\mu}_*(t) = \max_{i\in[K]} \widetilde{\mu}_i(t)$. Furthermore, let $\delta > 0$ be sufficiently small and

$$\begin{aligned}
\mathcal{A}(t) &:= \{\widetilde{\mu}_*(t) < \mu_1(t) - \delta\} \\
\mathcal{B}(t) &:= \{\widehat{\mu}_i(t) > \mu_i(t) + \delta\} \\
\mathcal{C}(t) &:= \left\{ N_i(t) < \frac{\log T}{d(\mu_i(t)+\delta, \mu_1 - \delta)} \right\}.
\end{aligned}$$

Event $\mathcal{A}(t)$ is related to the underestimation of the optimal arm, event $\mathcal{B}(t)$ is related to the overestimation of arm i, and event $\mathcal{C}(t)$ corresponds to the case that arm i is not sufficiently sampled. If none of these three events occurs, it is very unlikely that arm i is sampled. First, we decompose the number of draws of each arm as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i]\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{A}(t)]\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{B}(t)]\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{C}(t)]\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{A}^c(t), \mathcal{B}^c(t), \mathcal{C}^c(t)]\right] \\
&\leq \frac{\log T}{d(\mu_i(t)+\delta, \mu_1 - \delta)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{A}(t)]\right]}_{\text{(X)}} \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{B}(t)]\right]}_{\text{(Y)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{A}^c(t), \mathcal{B}^c(t), \mathcal{C}^c(t)]\right]}_{\text{(Z)}}.
\end{aligned}$$

In the following, we bound terms (X), (Y), and (Z) separately.

Bounding term (X):

Term (X), which implies the underestimation of the optimal arm by δ, is $O(\mathrm{poly}(1/\delta)) = O(1)$ as a function of T. Interestingly, bounding this non-leading term is the hardest part

in the analysis of TS. In bounding the overestimation of a suboptimal arm, we have

$$\begin{aligned}\sum_{t=1}^{T}\mathbf{1}[I(t)=i,\mathcal{A}(t)] &= \sum_{n=0}^{T}\sum_{t=1}^{T}\mathbf{1}[I(t)=i,\mathcal{A}(t),N_1(T)=n]\\ &= \sum_{n=0}^{T}\sum_{m=1}^{T}\mathbf{1}\left[m\leq\sum_{t=1}^{T}\mathbf{1}[I(t)=i,\mathcal{A}(t),N_1(T)=n]\right]\end{aligned}$$

Note that event

$$\left\{m\leq\sum_{t=1}^{T}\mathbf{1}[I(t)=i,\mathcal{A}(t),N_1(T)=n]\right\}$$

implies that $\widetilde{\mu}_1(t)\leq\widetilde{\mu}_*(t)\leq\mu_1-\delta$ occurred for the first m elements of $\{t:\mathcal{A}^c(t),N_1(T)=n\}$. Therefore,

$$\begin{aligned}&\mathbb{P}\left[m\leq\sum_{t=1}^{T}\mathbf{1}[I(t)=i,\mathcal{A}(t),N_1(T)=n]\middle|\widehat{\mu}_{1,n}\right]\\ &\leq(F^{\text{beta}}_{n\widehat{\mu}_{1,n}+1,n(1-\widehat{\mu}_{1,n})+1}(\mu_1-\delta))^m,\end{aligned}$$

where $\widehat{\mu}_{i,n}$ is the empirical mean reward of arm i with n samples. Accordingly, we get

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}[I(t)=i,\mathcal{A}(t)]\right] &\leq \sum_{n=0}^{T}\sum_{m=1}^{T}\mathbb{P}\left[m\leq\sum_{t=1}^{T}\mathbf{1}[I(t)=i,\mathcal{A}(t),N_1(T)=n]\right]\\ &\leq \sum_{n=0}^{T}\sum_{m=1}^{T}\mathbb{E}\left[(F^{\text{beta}}_{n\widehat{\mu}_{1,n}+1,n(1-\widehat{\mu}_{1,n})+1}(\mu_1-\delta))^m\right]\\ &\leq \sum_{n=0}^{T}\mathbb{E}\left[\left(\frac{1}{1-F^{\text{beta}}_{n\widehat{\mu}_{1,n}+1,n(1-\widehat{\mu}_{1,n})+1}(\mu_1-\delta)}-1\right)\right],\end{aligned}\tag{3.9}$$

where the last two expectations are taken with respect to $\widehat{\mu}_{1,n}$. Bounding (3.9) is rather technical. The existing papers [Kaufmann et al., 2012, Agrawal and Goyal, 2013a] give estimates for the partial Binomial sums [Jerábek, 2004]. In fact, one can show that (3.9) $= O(1)$. In general, bounding this term requires a special technique for each class of reward distributions. Besides the Bernoulli distribution, bounds on term (X) have been formulated for the Normal distribution [Honda and Takemura, 2014] and the one-parameter exponential distribution [Korda et al., 2013].

Bounding term (Y):

This term is related to the deviation of the empirical mean, and thus, it will have a large deviation inequality. Note that event $\{I(t)=i,\widehat{\mu}_i(t)>\mu_i+\delta,N_i(t)=n\}$ occurs at most

once and $\widehat{\mu}_i(t)$ is fixed during $N_i(t) = n$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t) = i, \mathcal{B}(t)]\right] &= \sum_{t=1}^{T} \mathbb{P}\{I(t) = i, \widehat{\mu}_i(t) > \mu_i + \delta\} \\ &= 1 + \sum_{n=1}^{T} \mathbb{P}\{\widehat{\mu}_{i,n} > \mu_i + \delta\} \\ &= 1 + \sum_{n=1}^{T} \mathrm{e}^{-2\delta^2 n} \quad \text{(by Hoeffding's inequality)} \\ &= O(1).\end{aligned}$$

Bounding term (Z):
Under $\{\mathcal{A}^c(t), \mathcal{B}^c(t), \mathcal{C}^c(t)\}$, it suffices to bound $\widetilde{\mu}_i(t)$, which is drawn from Beta$(N_i(t)\widehat{\mu}_i(t)+1, N_i(t)(1-\widehat{\mu}_i(t))+1)$. This term is specific to TS: in TS, one needs to bound the tail probability of the conjugate posterior distribution. Namely,

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t) = i, \mathcal{A}^c(t), \mathcal{B}^c(t), \mathcal{C}^c(t)]\right] \\ &= \sum_{t=1}^{T} \mathbb{P}[I(t) = i, \mathcal{A}^c(t), \mathcal{B}_i^c(t), \mathcal{C}_i^c(t)] \\ &\leq \sum_{t=1}^{T} \mathbb{P}\left[\widetilde{\mu}_i(t) > \mu_1(t) - \delta, \widehat{\mu}_i(t) \leq \mu_i + \delta, N_i(t) \geq \frac{\log T}{d(\mu_i(t) + \delta, \mu_1 - \delta)}\right] \\ &\leq \sum_{t=1}^{T} \frac{1}{T} = 1.\end{aligned}$$

where the last inequality is derived by using the Beta-Binomial transformation and the Chernoff inequality (c.f., (B.2) in Agrawal and Goyal [2013a]).
Final step:
From the continuity of the KL divergence, one can find $c_i = c_i(\mu_i, \mu_1)$ such that

$$d(\mu_i(t) + \delta, \mu_1 - \delta) \geq (1 - c_i\delta) d(\mu_i(t), \mu_1),$$

and thus, by letting $\delta = \epsilon/c_i$ and using the results on for terms (X), (Y), and (Z), we yield get (3.8). □

3.5 Deterministic Minimum Empirical Divergence (DMED)

DMED (deterministic minimum empirical divergence) proposed by Honda and Takemura [2010] is a relatively new algorithm for solving the stochastic bandit problem, which was proposed by Honda and Takemura [2010]. There is also a probabilistic version of the algorithm, called MED [Honda and Takemura, 2011].

Algorithm 6 DMED

Input: # of arms K.

$L_C, L_R \leftarrow [K], L_N \leftarrow \emptyset$.

Select each arm once and receive the rewards. $t \leftarrow K+1$.

while $t \leq T$ **do**

for $i \in L_C$ in an arbitrary fixed order, **do**

Select arm i and receive the corresponding reward. $t \leftarrow t+1$.

$L_R \leftarrow L_R \setminus \{i\}$.

$L_N \leftarrow L_N \cup \{j\}$ (without a duplicate) for all $j \notin L_R$ such that $\mathcal{J}_j(t)$ holds, where

$$\mathcal{J}_i(t) := \{N_i(t) d(\widehat{\mu}_i, \max_{i'} \widehat{\mu}_{i'}) \leq \log(t/N_i(t))\} \tag{3.10}$$

end for

L_C, $L_R \leftarrow L_N$, $L_N \leftarrow \emptyset$.

end while

One of the novel contributions of DMED to the understanding of the bandit problem is the introduction of the notion of "likelihood". The quantity $N_i(t) d(\widehat{\mu}_i, \max_{i'} \widehat{\mu}_{i'})$ in (3.10), which corresponds to the empirical divergence of each arm, can be considered as the negative log-likelihood that the arm is the optimal one. DMED continues sampling each arm until its empirical divergence becomes larger than $\log(t/N_i(t))$, which turns out to be sufficient for minimizing the expected regret. The computation of (3.10) is easy for Bernoulli rewards. One of the advantages of DMED lies in its extendability: this inequality can be efficiently computed for a large class of distributions with finite support. Although DMED was conceived from the Bayesian viewpoint, it requires no prior distribution.

The following theorem states that DMED is an optimal algorithm.

Theorem 10. *The regret of DMED is upper bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \neq 1} \frac{\Delta_i \log T}{d(\mu_i, \mu_1)} + o(\log T),$$

3.5.1 Analysis of DMED

In this section, we prove the following statement: for any suboptimal arm $i \neq 1$ and sufficiently small $\epsilon > 0$, we have

$$\mathbb{E}[N_i(T+1)] \leq (1+\epsilon) \frac{\log T}{d(\mu_i, \mu_1)} + O(1),$$

from which Theorem 10 easily follows. During In the proof, we will use the two following properties: namely,

- **(Property 1)** $\sum_{t=1}^{T} \mathbf{1}\{I(t)=i\} \leq \sum_{t=1}^{T} \mathbf{1}\{\mathcal{J}_i(t)\} + 2$. Moreover,
- **(Property 2)** If $\mathcal{J}_i(t)$ holds, then the arm will be drawn within K rounds.

The proof here follows the essentially the same steps to as the one in Honda and Takemura [2010], with some improvements made on to the arguments.

Let $\widehat{\mu}_*(t) = \max_{i\in[K]} \widehat{\mu}_i(t)$ and $\delta > 0$ be such that $\mu_2 + \delta < \mu_1$. Let

$$\begin{aligned}
\mathcal{A}(t) &:= \{\widehat{\mu}_*(t) \leq \mu_2 + \delta\} \\
\mathcal{B}(t) &:= \{\widehat{\mu}_1 \leq \mu_1 - \delta\} \\
\mathcal{C}(t) &:= \{N_i(t) \leq \frac{(1+\epsilon)\log T}{d(\mu_i, \mu_1)}\}.
\end{aligned}$$

Event $\mathcal{A}(t)$ is related to the underestimation of the optimal arm, event $\{\mathcal{A}^c(t), \mathcal{B}(t)\}$ is related to the overestimation of suboptimal arms, and $\mathcal{C}(t)$ corresponds to the case that arm i is not sufficiently sampled. We first decompose the number of draws on each arm as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i]\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[\mathcal{A}(t)]\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[\mathcal{A}^c(t), \mathcal{B}(t)]\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{C}(t)]\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{A}^c(t), \mathcal{B}^c(t), \mathcal{C}^c(t)]\right] \\
&\leq \frac{(1+\epsilon)\log T}{d(\mu_i, \mu_1)} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[\mathcal{A}(t)]\right]}_{\text{(X)}} \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[\mathcal{A}^c(t), \mathcal{B}(t)]\right]}_{\text{(Y)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{B}^c(t), \mathcal{C}^c(t)]\right]}_{\text{(Z)}}.
\end{aligned}$$

In the following, we bound terms (X), (Y), and (Z) separately.

Bounding term (X):

Term (X) is related to the underestimation of the optimal arm. Recall that, in bounding the underestimation of the optimal arm, TS requires an elaborate technique which, which is highly specific to the Bernoulli distribution. KL-UCB also requires the self-normalizing bound to this end. On the contrary, the following bound on term (X) does not require any special technique.

We have

$$\sum_{t=1}^{T} \mathbf{1}[\mathcal{A}(t)] = \sum_{t=1}^{T} \mathbf{1}\{I(t)=i, \widehat{\mu}_1(t) \leq \mu_2 + \delta, \widehat{\mu}_*(t) \leq \mu_2 + \delta\}.$$

Under event $\{\widehat{\mu}_*(t) \le \mu_2 + \delta\}$, if

$$t/N_i(t) \ge \exp\left(N_i(t) d(\widehat{\mu}_1(t), \mu_2 + \delta)\right)$$

then $\mathcal{J}_1(t)$ holds and arm 1 will be drawn within K rounds (by Property 2). Therefore,

$$\sum_{t=1}^{T} \mathbf{1}\left[\widehat{\mu}_1(t) \le \mu_2 + \delta, N_1(t) = n, \widehat{\mu}_*(t) \le \mu_2 + \delta\right] \le n \exp\left(n d(\widehat{\mu}_{1,n}, \mu_2 + \delta)\right) + K,$$

where $\widehat{\mu}_{i,n}$ be the empirical mean reward of arm i with n samples. Let $P_1(x) = \mathbb{P}[\widehat{\mu}_{1,n} \le \mu_2 + \delta, d(\widehat{\mu}_{1,n}, \mu_2 + \delta) \ge x]$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left[\widehat{\mu}_1(t) \le \mu_2 + \delta, N_1(t) = n, \widehat{\mu}_*(t) \le \mu_2 + \delta\right]\right] \\
&\le \int_0^{\log(1/(1-\mu_2-\delta))} (n\mathrm{e}^{nx} + K)\, d(-P_1(x)) \\
&\le K P_1(0) + \int_0^{\log(1/(1-\mu_2-\delta))} n\mathrm{e}^{nx} d(-P_1(x)) \\
&\le K P_1(0) + \left[n\mathrm{e}^{nx}(-P_1(x))\right]_{x=0}^{\log(1/(1-\mu_2-\delta))} + \int_0^{\log(1/(1-\mu_2-\delta))} n^2 \mathrm{e}^{nx} P_1(x) dx \\
&\qquad \text{(by integration by parts)} \\
&\le (n+K) P_1(0) + \int_0^{\log(1/(1-\mu_2-\delta))} n^2 \mathrm{e}^{n d(\mu_x, \mu_2+\delta)} \mathrm{e}^{-n d(\mu_x, \mu_1)} dx,
\end{aligned}$$

where $\mu_x \le \mu_2 + \delta$ is such that $d(\mu_x, \mu_2 + \delta) = x$. By using Fact 36, we have

$$\begin{aligned}
&\int_0^{\log(1/(1-\mu_2-\delta))} n^2 \mathrm{e}^{n d(\mu_x, \mu_2+\delta)} \mathrm{e}^{-n d(\mu_x, \mu_1)} dx \\
&\le \int_0^{\log(1/(1-\mu_2-\delta))} n^2 \mathrm{e}^{-n C_1(\mu_1, \mu_2+\delta)} dx \\
&\le \log\left(1/(1-\mu_2-\delta)\right) n^2 \mathrm{e}^{-n C_1(\mu_1, \mu_2+\delta)},
\end{aligned}$$

where $C_1(\mu, \mu_2) = (\mu - \mu_2)^2/(2\mu(1-\mu_2))$. Summing these results over n yields

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left[\widehat{\mu}_1(t) \le \mu_2 + \delta, \widehat{\mu}_*(t) \le \mu_2 + \delta\right]\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \sum_{n=1}^{\infty} \mathbf{1}\left[\widehat{\mu}_1(t) \le \mu_2 + \delta, \widehat{\mu}_*(t) \le \mu_2 + \delta, N_i(t) = n\right]\right] \\
&\le \sum_{n=1}^{\infty} \left((n+K) P_1(0) + \log\left(1/(1-\mu_2-\delta)\right) n^2 \mathrm{e}^{-n C_1(\mu_1, \mu_2+\delta)}\right) \\
&\le \sum_{n=1}^{\infty} \left((n+K) \mathrm{e}^{-n d(\mu_2+\delta, \mu_1)} + \log\left(1/(1-\mu_2-\delta)\right) n^2 \mathrm{e}^{-n C_1(\mu_1, \mu_2+\delta)}\right) \\
&= \left(\frac{\mathrm{e}^{d(\mu_2+\delta, \mu_1)}}{(\mathrm{e}^{d(\mu_2+\delta, \mu_1)} + 1)^2} + K \frac{1}{\mathrm{e}^{d(\mu_2+\delta, \mu_1)} + 1} + \log\left(\frac{1}{1-\mu_2-\delta}\right) \frac{\mathrm{e}^{C_1(\mu_1, \mu_2+\delta)} (\mathrm{e}^{2C_1(\mu_1, \mu_2+\delta)} - 1)}{(\mathrm{e}^{C_1(\mu_1, \mu_2+\delta)} + 1)^4}\right) \\
&< +\infty,
\end{aligned}$$

where we have used the fact that $\sum_{n=1}^{\infty} \mathrm{e}^{-nx} = 1/(\mathrm{e}^x + 1)$, $\sum_{n=1}^{\infty} n\mathrm{e}^{-nx} = \mathrm{e}^x/(\mathrm{e}^x+1)^2$, and $\sum_{n=1}^{\infty} n^2\mathrm{e}^{-nx} = \mathrm{e}^x(\mathrm{e}^{2x} - \mathrm{e}^x)/((\mathrm{e}^x+1)^4)$. In summary, term (X) is $O(1)$ as a function of T.

Bounding term (Y):

Note that

$$\sum_{t=1}^{T} \mathbf{1}[\mathcal{A}^c(t), \mathcal{B}(t)] = \sum_{t=1}^{T} \mathbf{1}[\widehat{\mu}_*(t) > \mu_2 + \delta, \widehat{\mu}_1 \leq \mu_1 - \delta]$$

implies $\cup_{i\in[K]}\{\widehat{\mu}_i(t) = \widehat{\mu}_*(t), |\widehat{\mu}_i(t) - \mu_i| \geq \delta\}$. Here,

$$\sum_{t=1}^{T} \mathbf{1}[\widehat{\mu}_i(t) = \widehat{\mu}_*(t), |\widehat{\mu}_i(t) - \mu_i| \geq \delta] = \sum_{t=1}^{T}\sum_{n=1}^{T} \mathbf{1}[\widehat{\mu}_i(t) = \widehat{\mu}_*(t), |\widehat{\mu}_i(t) - \mu_i| \geq \delta, N_i(t) = n].$$

Suppose that $\{\widehat{\mu}_i(t_0) = \widehat{\mu}_*(t_0), N_i(t_0) = n\}$ occurs the first time at round t_0. Then, $\mathcal{J}_i(t)$ holds, and arm i is drawn within K rounds (from Property 2). Therefore, $\mathbf{1}[\widehat{\mu}_i(t) = \widehat{\mu}_*(t), |\widehat{\mu}_i(t) - \mu_i| \geq \delta, N_i(t) = n]$ occurs at most K rounds. By using this fact, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[\widehat{\mu}_i(t) = \widehat{\mu}_*(t), |\widehat{\mu}_i(t) - \mu_i| \geq \delta]\right] &\leq K\mathbb{E}\left[\sum_{n=1}^{T} \mathbf{1}[|\widehat{\mu}_{i,n}(t) - \mu_i| \geq \delta]\right] \\
&\leq 2K\sum_{n=1}^{T} \mathrm{e}^{-2n\delta^2} \quad \text{(by Hoeffding's inequality)} \\
&= \frac{2K}{\mathrm{e}^{\delta^2} - 1} < \infty.
\end{aligned}$$

In summary, term (Y) is $O(1)$.

Bounding term (Z):

This term is related the overestimation of arm i after a sufficient number of samples. With a sufficient number of samples, the deviation of $\widehat{\mu}_i(t)$ from μ_i is likely to be small.

Therefore, this term is expected to be $O(1)$.

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{B}^c(t), \mathcal{C}^c(t)] \\
&\sum_{t=1}^{T} \mathbf{1}\left[I(t)=i, \widehat{\mu}_* > \mu_1 - \delta, N_i(t) \geq \frac{(1+\epsilon)\log T}{d(\mu_i,\mu_1)}\right] \\
&\leq \sum_{t=1}^{T} \mathbf{1}\left[I(t)=i, \widehat{\mu}_* > \mu_1 - \delta, N_i(t) \geq \frac{(1+\epsilon)\log T}{d(\mu_i,\mu_1)}\right] \\
&\leq 2 + \sum_{t=1}^{T} \mathbf{1}\left[\mathcal{J}_i(t), \widehat{\mu}_* > \mu_1 - \delta, N_i(t) \geq \frac{(1+\epsilon)\log T}{d(\mu_i,\mu_1)}\right] \\
&\qquad \text{(from Property 1)} \\
&\leq 2 + \sum_{n=\lceil \frac{(1+\epsilon)\log T}{d(\mu_i,\mu_1)} \rceil}^{T} \mathbf{1}\left[\bigcup_{t=1}^{T} \{n d(\widehat{\mu}_{i,n}, \mu_1 - \delta) \leq \log t\}\right] \\
&\leq 2 + \sum_{n=\lceil \frac{(1+\epsilon)\log T}{d(\mu_i,\mu_1)} \rceil}^{T} \mathbf{1}\left[\frac{(1+\epsilon)\log T}{d(\mu_i,\mu_1)} d(\widehat{\mu}_{i,n}, \mu_1 - \delta) \leq \log T\right] \\
&\leq 2 + \sum_{n=\lceil \frac{(1+\epsilon)\log T}{d(\mu_i,\mu_1)} \rceil}^{T} \mathbf{1}\left[d(\widehat{\mu}_{i,n}, \mu_1 - \delta) \leq \frac{d(\mu_i,\mu_1)}{1+\epsilon}\right].
\end{aligned}
$$

If δ is sufficiently small, there exists a sufficiently small $c = c(\epsilon, \delta) > 0$ such that

$$
\mathbb{P}\left[d(\widehat{\mu}_{i,n}, \mu_1 - \delta) \leq \frac{d(\mu_i,\mu_1)}{1+\epsilon}\right] \leq \mathrm{e}^{-nc}. \tag{3.11}
$$

In summary, we have

$$
\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}[I(t)=i, \mathcal{B}^c(t), \mathcal{C}^c(t)]\right] \leq 2 + \sum_{n=1}^{\infty} \mathrm{e}^{-nc} = O(1). \tag{3.12}
$$

Summing up the bounds on terms (X), (Y), and (Z) completes the proof.

3.6 Performance of the Algorithms

We examined the performances of the bandit algorithms in a computer simulation[*3]. The simulation involved ten Bernoulli arms with $\mu_1 = 0.1, \mu_2, \ldots, \mu_4 = 0.05, \mu_5, \ldots, \mu_7 = 0.02, \mu_8, \ldots, \mu_{10} = 0.01$. The algorithms were as follows: ϵ_t-greedy [Auer et al., 2002a] with $\epsilon_t = 0.1/t$, UCB1 (Algorithm 3) with $\alpha = 1$, MOSS [Audibert and Bubeck, 2009], and UCB-V [Audibert et al., 2009]. KL-UCB (T) is Algorithm 4 with $c = 0$, and KL-UCB (ToN) is a version of KL-UCB with its $\log t$ factor replaced by $\log(t/N_i(t))$, which

[*3] The simulation was based on the author's open source software, which is available at URL: https://github.com/jkomiyama/banditlib.

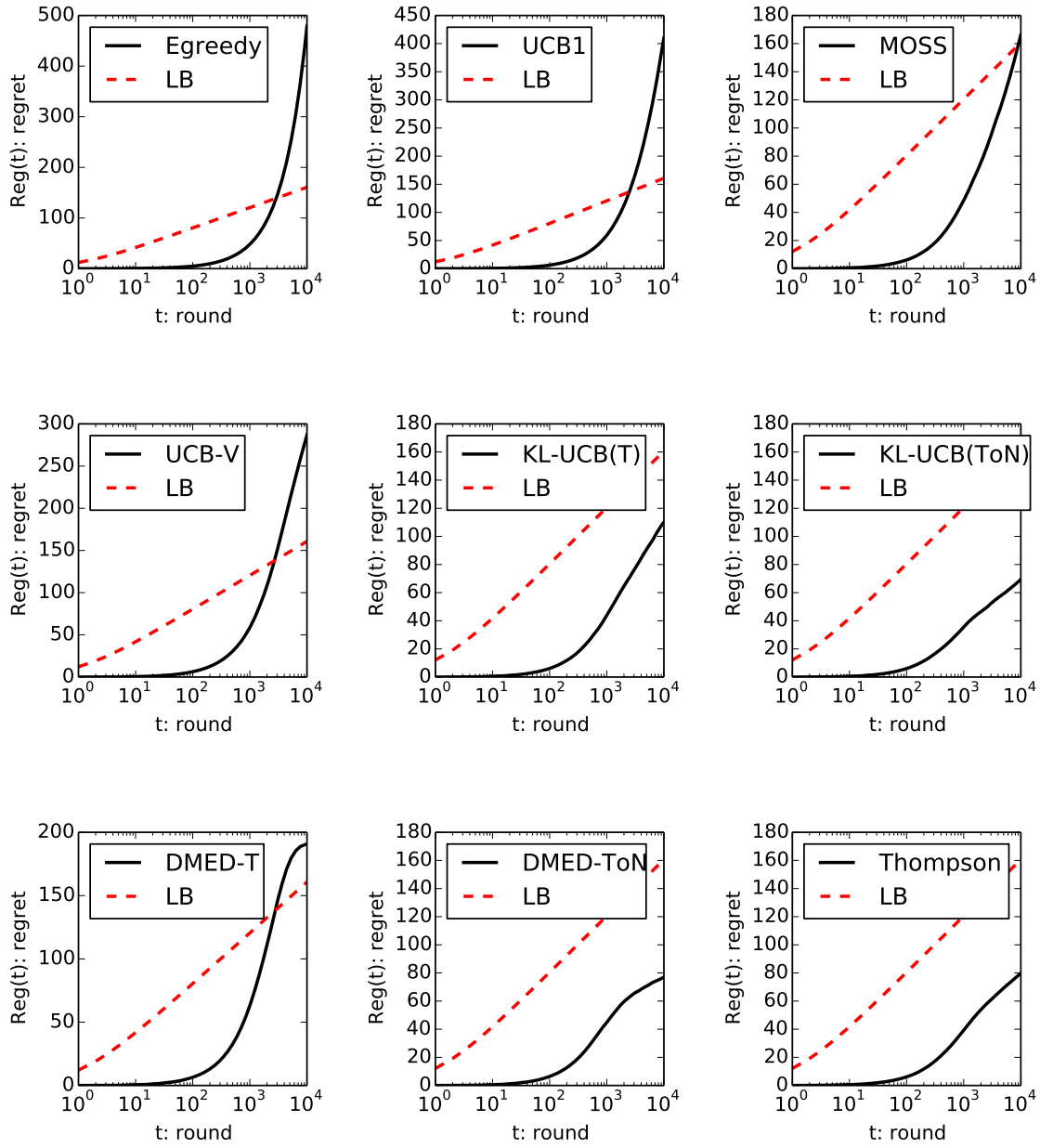


Fig. 3.3. Regret-round semilog plots of algorithms. LB is the asymptotic lower bound ($= \sum_{i \neq 1} \frac{\Delta_i \log t}{d(\mu_i, \mu_1)}$). The results are averaged over 10,000 trials.

is referred to as KL-UCB+ in Garivier and Cappé [2011]. DMED-ToN is Algorithm 6, and DMED-T is a version with $\log(t/N_i(t))$ in (3.10) replaced by $\log t$. TS is Algorithm 5. Note that the KL-UCBs, DMEDs, and TS are asymptotically optimal in the sense that their regret bound asymptotically matches the lower bound, whereas the other algorithms are not. The results are shown in Figure 3.3. KL-UCB (ToN), DMED-ToN, and TS performed the best and were very close to each other, which is probably related to the

fact that they are based on the optimal exploration rate, as discussed in Section 3.2.

3.7 Discussion

Here, we examine a number of topics related to the bandit algorithms. Section 3.7.1 discusses the asymptotic and finite-time properties of the analyses, and section 3.7.2 looks at the distribution-independent regret, which is yet another regret that holds for any model parameter $\{\mu_i\}$. Section 3.7.3 examines the Bayes risk of the undiscounted cumulative reward, and section 3.7.4 describes the relation between the space of the parameters and the regret lower bound. Section 3.7.5 discusses the extendability of the bandit algorithms.

3.7.1 Asymptotic and finite-time analyses

We have demonstrated that KL-UCB, TS, and DMED have asymptotic regret bounds. Recall that an algorithm is called asymptotically optimal if it has a regret bound that implies

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \neq 1} \frac{\Delta_i \log T}{d(\mu_i, \mu_1)} + o(\log T). \tag{3.13}$$

Note that the small-o notation in (3.13) only specifies an asymptotic property: an $o(\log T)$ function is arbitrarily smaller than $\log T$ given sufficiently large T, but it does not specify how fast it becomes bounded. On the other hand, a finite-time analysis, which gives an explicit regret bound for given T, and thus guarantees the performance of the algorithm in a more explicit way, is important in the field of machine learning.

The properties of algorithms we consider in this section are illustrated in Figure 3.4. Here, we will not explicitly discuss the finite-time property of the regret bounds because it is not difficult to obtain a finite-time bound for the Bernoulli rewards. In general, whether a finite-time analysis is possible or not depends on the distribution of the rewards.

3.7.2 Distribution-independent regret bound

The regret lower bound is described in terms of the KL divergence between the distributions (Section 2.3.2). This bound is asymptotic for $T \to \infty$ when we regard the model parameters $\{\mu_i\}_{i \in [K]}$ as constants. In this sense, it is distribution-dependent. In the multi-armed bandit problem, another kind of regret bound is known: namely, a distribution-independent regret bound that holds for any parameter $\{\mu_i\}_{i \in [K]}$. It is easy to show that an algorithm with an $O(K \log T / \Delta_2)$ distribution-dependent regret bound (e.g., UCB1, KL-UCB, TS, and DMED) also has an $O(\sqrt{TK \log T})$ distribution-free regret (e.g., Theorem 2.2 in Bubeck [2010] or Theorem 2 in Agrawal and Goyal [2013a]). On the other hand, the distribution-free regret is lower-bounded by $\Omega(\sqrt{KT})$.

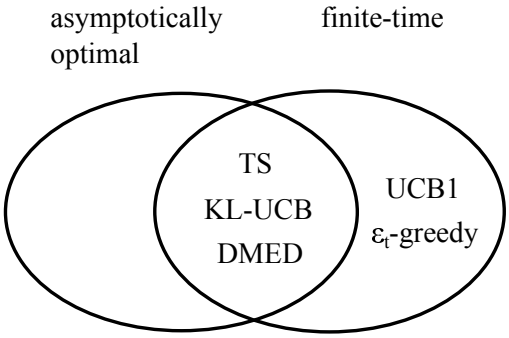


Fig. 3.4. Asymptotic and finite-time properties of the bandit algorithms that we discuss in this thesis. Asymptotical optimality means that an algorithm is associated with a regret bound that satisfies inequality (3.13). The finite-time property indicates the availability of an explicit formula of the regret bound for given T. Note that the regret of all algorithms considered in this section is finite-time for Bernoulli rewards.

Theorem 11. (Appendix A in Auer et al. [2002b]) *For any algorithm, there exists a set of parameters* $\{\mu_i\}_{i\in[K]}$ *such that the regret is lower bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \frac{1}{20}\sqrt{KT}.$$

There is a logarithmic gap between the $O(\sqrt{TK\log T})$ regret upper bound and the $\Omega(\sqrt{KT})$ regret lower bound of Theorem 11. The MOSS algorithm [Audibert and Bubeck, 2009] is a stochastic bandit algorithm with $O(\sqrt{KT})$ regret, and it fills the logarithmic gap. To the best of our knowledge, the optimal leading constant of $O(\sqrt{KT})$ regret is unknown.

3.7.3 Bayesian view of regret

Unlike the Bayesian approach that we explained in Section 2.2, in this section, we focus on the case in which the rewards are undiscounted. Lai [1987] considered a maximization of the undiscounted cumulative reward from the Bayesian point of view. They also proposed an algorithm that is similar to KL-UCB+ [Garivier and Cappé, 2011], which is asymptotically optimal in the sense of the Bayes risk. Interestingly, to show the asymptotic optimality of the algorithm, they first prove the optimality of KL-UCB+ in the frequentist sense (i.e., as a stochastic bandit algorithm), and after that, they prove the optimality of the algorithm with respect to the Bayes risk:

$$\mathbb{E}_\Pi\left[T\max_i \mu_i - \sum_{t=1}^{T}\widehat{X}_{I(t)}(t)\right],$$

where the expectation is taken over a given prior distribution Π. Although the optimality of a stochastic bandit algorithm does not imply the optimality of a Bayesian bandit algorithm [Kaufmann, 2014], the frequentist and Bayesian criteria are rather similar when we adopt undiscounted reward.

3.7.4 Model class and regret

Suppose that in the stochastic bandit problem a forecaster is informed of the fact $\mu_1 > \mu_2$. In this case, how much exploration over the arm 2 does the forecaster need to do? Not a single draw of arm 2 is required, since the forecaster knows arm 2 is not optimal. As demonstrated in this example, the optimal regret depends on the space of the model parameters under consideration. Given an additional restriction on the model parameter space, the regret can be made smaller than that of searching the entire model space.

Graves and Lai [1997] generalized the multi-armed bandit problem so that the model parameters can be shared among arms[*4]. Let $\theta \in \Theta$ be a set of unknown model parameters and Θ be compact. Let $P_i(\theta)$ be a distribution parameterized by θ. At each round, the algorithm draws an arm $I(t)$ and receives a corresponding reward $\widehat{X}_{I(t)} \sim P_{I(t)}(\theta)$. To simplify the discussion, we will assume that $P_i(\theta)$ is Bernoulli and $\mu_i = \mu_i(\theta)$ is the expectation of the arm i with parameter θ. Moreover, we will assume that arm 1 is optimal (i.e., the unique maximizer of the expectation). The goal of the algorithm is to minimize the regret,

$$\mathrm{Reg}(T) := T\mu_1(\theta) - \sum_{t=1}^{T} \mu_{I(t)}(\theta).$$

Under some mild assumptions on the reward distributions, the regret of any strongly consistent algorithm can be asymptotically lower bounded as follows. As in the standard bandit problem, an strongly consistent algorithm needs to confirm that each arm $i \neq 1$ is suboptimal. Let

$$\mathcal{R}_1 := \left\{ \{c_j\} \in [0,\infty)^{K-1} : \inf_{\theta' \in \Theta : \arg\max_{i \in [K]} \mu_i(\theta') \neq 1, \mu_1(\theta') = \mu_1(\theta)} \sum_{j \neq 1} c_j d(\mu_j(\theta), \mu_j(\theta')) \geq 1 \right\}.$$

Intuitively, receiving the rewards of each suboptimal arm i for $c_i \log t$ times is sufficient for making sure that the optimal arm is not 1 with a confidence level $1/t$. Thus,

$$\mathbb{E}[\mathrm{Reg}(T)] \geq C_* \log T - o(T),$$

where $C_* = \inf_{\{c_j\} \in \mathcal{R}_1} \sum_{j \neq 1} c_j (\mu_1(\theta) - \mu_j(\theta))$. The quantity $C_* \log T$ defines the minimum amount of exploration to ensure that each arm $i \neq 1$ is suboptimal with a confidence

[*4] They also extended the bandit problem to involve a state (i.e., Markovian), which we will not discuss here.

level $1/t$. A more general version of this result is stated in Theorem 1 in Graves and Lai [1997]. The regret lower bound of the standard multi-armed bandit problem that we discussed in Section 2.3.2 can be considered a special case where the model parameters of the arms are disjoint: $\theta = (\mu_1, \mu_2, \ldots, \mu_K) \in \Theta = (0,1)^K$. One can easily confirm that $C_* = \sum_{i \neq 1} \Delta_i / d(\mu_i, \mu_1)$.

By using the above bound, one can derive the lower bound on the regret for the unimodal bandit problem [Yu and Mannor, 2011], a version of the stochastic bandit problem. In the unimodal bandit problem, arms are associated with an undirected graph. The expected reward of arms must be unimodal on the graph: in this sense, the model parameter is restricted. This is an example in which the model parameter space is smaller than the entire space. Because of the smaller model space, the regret of the unimodal bandits is smaller than that of the standard multi-armed bandit [Combes and Proutiere, 2014].

3.7.5 Extending the bandit algorithms

In this section, we consider the applicability of the aforementioned algorithms in solving extensions to the bandit problem. Since UCB is the most widely used bandit algorithm, we will consider it first. Recall that at each round, UCB selects the arm with the largest upper confidence bound. The behavior of the algorithm is illustrated in Figure 3.5. For large t, the upper confidence bound of arm 1 is close to μ_1. Each suboptimal arm is sampled until its upper confidence bound reaches μ_1. UCB provides an efficient solution as long as (i) each arm is associated with its expected reward, and (ii) exploring each arm with a large uncertainty is an efficient search method. These criteria are satisfied by not only the multi-armed bandit problem but also by certain classes of continuous bandit problems (Section 7.2.1) in which the number of arms is infinite. In these continuous bandit problems, a metric is associated with the arms and arms nearby each other on the metric are likely to have similar expectations. In these problems, exploring an arm with a large uncertainty reduces the uncertainty around the arm and is thus an effective way of determining the optimal arm.

Nevertheless, UCB is not always an efficient way of solving extensions of the bandit problem. The dueling bandit problem, which we discuss in Chapter 6, is one such case. In the dueling bandit problem, the feedback is limited to a relative comparison between arms. Since the expectation of each arm is not directly observable in this problem, UCB is not directly applicable. Allocating a confidence bound on each pair of arms is possible, but it turns out to be not very effective: comparing each pair until its upper confidence bound shrinks results in an $O(K^2 \log T)$ regret since the number of pairs is $O(K^2)$, whereas the optimal regret lower bound in the dueling bandit is $\Omega(K \log t)$ [Yue et al., 2012]. To circumvent this problem, in Chapter 6, we provide a version of the DMED algorithm and derive its optimal regret bound: we define the likelihood of each arm to be optimal

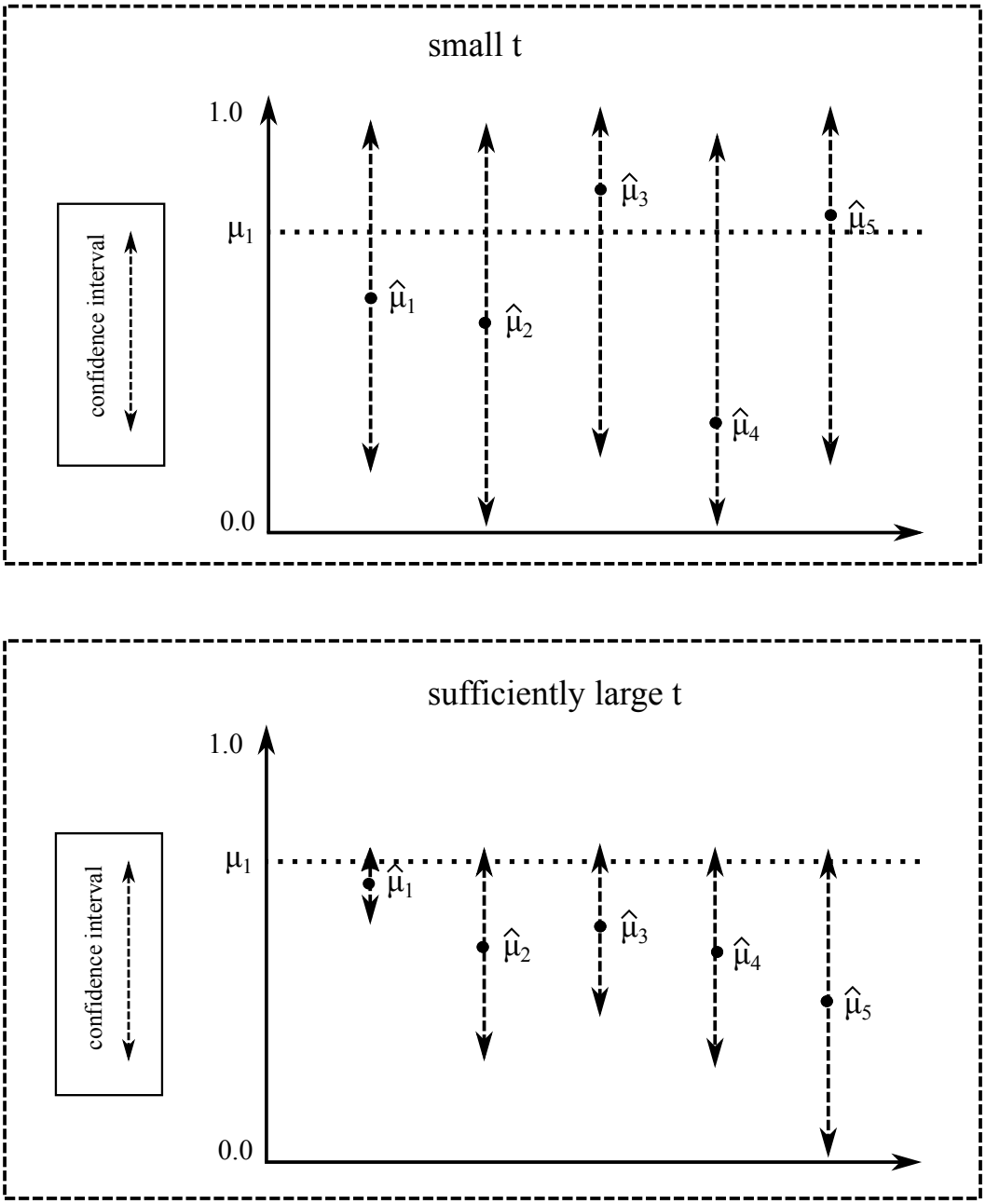


Fig. 3.5. Illustration of confidence bounds in UCB. The arms are not sufficiently explored after a small number of rounds t. Their confidence bound is very large, and some of their means are larger than those of the optimal ones. When t is sufficiently large, arm 1 (optimal arm) is sampled $O(t)$ times and its confidence bound is very small. The other arms are sampled until each of their upper confidence bounds is close to μ_1.

and explore each arm until its likelihood is sufficiently small, which requires $O(\log T)$ observations per arm. As a result, the proposed algorithm has $O(K \log T)$ regret.

Thompson sampling appears to be effective in most of the problems to which UCB

applies. In particular, TS has been shown to be effective on most of the continuous bandit problems. The author is not sure to what extent the idea of posterior sampling is applicable. Note also that there are some problems in which all of UCB, TS, and DMED are unsuccessful: namely, the many-armed bandit problems in which the number of arms is very large. Unlike the continuous-armed bandit problems, the metric structure of the arms is not available in many-armed bandit problems. Instead, one can exploit assumptions on the tail-probability on the sampled arm's expectation. We discuss these ideas in Section 7.2.2.

Chapter 4

Multi-armed Bandit Problem with Lock-up Periods

In this chapter, we investigate a version of the stochastic multi-armed bandit problem in which the forecaster's choice is restricted. In this version, rounds are divided into lock-up periods and the forecaster must select the same arm throughout a period. As explained in Chapter 3, there has been much work on finding optimal algorithms for the stochastic multi-armed bandit problem. However, their use under restricted conditions is not obvious. We extend the application ranges of these algorithms by proposing a natural conversion method from algorithms for the stochastic bandit problem to ones for the multi-armed bandit problem with lock-up periods. We prove that the regret of the converted algorithms is $O(\log T + L_{\max})$, where T is the total number of rounds and $L_{\max}$ is the maximum size of the lock-up periods. The regret is preferable, except for the case when the maximum size of the lock-up periods is large. For that case, we propose a meta-algorithm that results in a smaller regret by using an empirical optimal arm for large periods. We empirically compare and discuss these algorithms[*1]. The notation in this chapter is summarized in Table 4.1.

4.1 Motivation

In studying the bandit problems, the forecaster has the freedom to select an arbitrary arm at each round. However, in practical situations there are various restrictions for selecting arms. Many requirements, such as operation ease or resource constraints prevent the forecaster from free allocation. The examples below are typical scenarios.

Example 1. (A/B testings) A/B testing is a well-known method when releasing new web page features. By comparing the user responses for multiple versions of web pages, administrators can estimate the effectiveness of the releases. There are many targets of

[*1] The contents of this chapter were published in Komiyama et al. [2013a] and Komiyama et al. [2013b].

Table. 4.1. Notation used in Chapter 4.

$\mathbf{1}\{A\}$	:=	1 if A is true and 0 otherwise.
K	:=	Number of the arms.
T	:=	Number of the rounds.
N	:=	Number of the lock-up periods.
$I(t)$	:=	The arm that is selected in round t.
$\widehat{X}_i(t)$	:=	Reward of arm i at round t.
μ_i	:=	Expected reward of arm i.
μ_*	:=	$\max_i \mu_i$.
$\widehat{\mu}_i(t)$	:=	Empirical mean reward of arm i at round t.
i^*	:=	$\arg\max_i \mu_i$ (assumed to be unique).
Δ_i	:=	$\mu_* - \mu_i$.
Δ	:=	$\min_{i\in\{1,\dots,K\}, i\neq i^*} \Delta_i > 0$.
$T_i(t)$	:=	$\sum_{t'=1}^{t} \mathbf{1}\{I(t') = i\}$.
L_n	:=	Size of lock-up period n.
$L_{(j)}$	:=	Size of the j-th largest lock-up period.
$L_{\max}$	:=	Size of the largest lock-up period.
s_n, f_n	:=	Start and end of of the lock-up period n.

A/B testing, e.g., ad placements, emails and top pages. Optimizing user attention is of great importance for most large-scale websites. However, there are many constraints preventing optimal allocation. In a web system, the update of the database is delayed, and the click feedback of a user takes some time since it is displayed on the user's screen.

Example 2. (Clinical trials) Clinical trials are conducted in the final stages of drug development. The aim of such trials is to ensure the effectiveness and safety of newly developed drugs. There are many conditions necessary for this, e.g., amounts of drugs, placebo conditions, patients conditions. The trials are divided into many test phases. Between each test, the results of the previous test are reported. The next test is based on the information up to that and including that of the previous test. For the simplicity of operation, each test should be done with a single option. We would like to optimize the allocation even within these restrictions.

Essentially, these problems lie midway between sequential and batch problems. Forecasters are restricted to selecting the same option for certain rounds due to external constraints. Also, the sum of rewards is the quantity to optimize. To model these scenarios, we propose and study a multi-armed bandit problem with lock-up periods (lock-up bandit). The term "lock-up period" is a financial term meaning the predefined amount of time during which people concerned cannot sell shares. In the problem, we define the

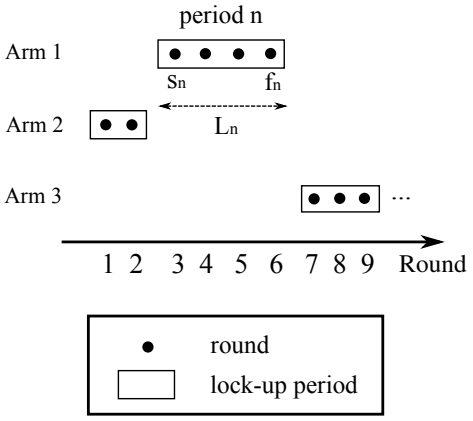


Fig. 4.1. Lock-up bandit. Black dots represent rounds and rectangles represent lock-up periods.

lock-up period as a set of successive rounds where the forecaster cannot change the arm to pull.

The rest of this chapter is structured as follows. In Section 4.2 we formalize the proposed problem and discuss related work. In the following sections, we start from the stochastic multi-armed bandit algorithms and prove they can keep small regrets in the existence of lock-up periods. The state-of-the-art algorithms for the stochastic bandit problem are not directly applicable to restricted environments. In Section 4.3, we discuss the natural conversion from the standard stochastic bandit algorithms to lock-up bandit ones. We prove the upper-bound regret of converted UCB, which is optimal up to a constant factor when the periods are small compared with the total number of rounds. The regrets of the converted algorithms are upper bounded by the size of the largest lock-up period. In some cases, there are large lock-up periods and in these cases the regret is linear to that much sizes. For such a case, we want to minimize the regret during these the large periods. In Section 4.4, we propose the balancing and recommendation (BaR) meta-algorithm, which effectively reduces the regret losses in large periods. The regret of this meta-algorithm is represented using the cumulative and simple regrets of the base algorithm. In Section 4.5, we show the results of two sets of experiments we conducted. The first was the empirical relation between the period size and the regret. The second set of experiments involved the before-after analysis of the BaR meta-algorithm. Finally, we conclude this chapter in Section 4.6.

4.2 Multi-armed Bandit Problem with Lock-up Periods

Our lock-up bandit is based on the stochastic bandit problem, in which the forecaster can select one arm at each round. However, in lock-up bandit, the rounds are divided into lock-up periods and the forecaster must select one arm for each lock-up period (Figure 4.1).

The lock-up bandit problem is formally defined as follows. There are K arms associated with constant reward distributions $\{P_1, ..., P_K\}$. We assume the supports of the reward distributions are in $[0, 1]$[*2]. There are T rounds that are divided into lock-up periods $L_1, ..., L_N$ where $\sum_{n=1}^{N} L_n = T$. We denote the start and end rounds of each periods as $(s_1, f_1), ..., (s_N, f_N)$. Note that $s_1 = 1$, $f_N = T$, $s_{n+1} - 1 = f_n$ and $L_n = f_n - s_n + 1$ hold for all periods $n \in [1, ..., N-1]$. Before the start of the first round, the forecaster is notified of K and $L_1, ..., L_N$. On each round $t = 1, ..., T$, if the round is the start of a period, the forecaster selects an arm. If not, he or she uses the same arm as the previous round. We denote the selected arm at round t as I_t. After selecting an arm, the forecaster receives the reward $\widehat{X}_{I(t)}(t) \sim P_{I_t}$.

The goal of the forecaster is to minimize the (cumulative) regret

$$\mathrm{Reg}(T) = \mu_* T - \sum_{i=1}^{K} \mu_i T_i(T),$$

where $\mu_i = \mathbb{E}[P_i]$, μ_* is $\max_i \mu_i$, and $T_i(T)$ is the number of rounds arm i was selected in T rounds. For the ease of discussion, we assume the optimal arm $i^* = \arg\max_i \mu_i$ is unique. We also use the gap $\Delta_i = \mu_* - \mu_i$ and the minimum nonzero gap $\Delta = \min_{i \in \{1,...,K\}, i \neq i^*} \Delta_i$. By the definition above, selecting suboptimal arm i increases the regret by Δ_i and that can be considered as a loss.

Remark 12. Multi-armed bandit with lock-up periods $L_1, ..., L_N, \sum_n L_n = T$ is more difficult than T-round stochastic bandit, where the forecaster can switch arms for every round. The consistency-based argument of the regret lower bound for the stochastic bandit problem also holds for the one in the lock-up bandit problem with the same number of rounds.

4.2.1 Round-wise notation and period-wise notation

Throughout this chapter, we use t as a variable representing a round and n as a variable representing a period. We use i as a variable representing an arm. For example, the number of rounds the arm i was selected in T rounds is denoted as follows.

$$T_i(T) = \sum_{t=1}^{T} \mathbf{1}\{I_t = i\},$$

where $\mathbf{1}\{A\}$ is 1 if A is true and 0 otherwise. As the forecaster must select one arm during a period, we can denote I_n to represent the arm selected in period n. Also, we use the notation $T_i(L_1, ..., L_N)$ for the number of draws to explicitly denote the lock-up periods

[*2] Generalization to any finite support $[a, b]$ with $a, b \in \mathbb{R}$ is easy.

$L_1, \ldots, L_N$. Namely,

$$T_i(L_1, ..., L_N) = \sum_{n=1}^{N} L_n \mathbf{1}\{I_n = i\}.$$

4.2.2 Period ordering

In lock-up bandit, the order of periods matters. Remember the length of first period is denoted as L_1 and that of the second is L_2 , etc. For example, the lock-up bandit problem with $L_1, ..., L_9 = 1, L_{10} = 10$ is much easier than the one with $L_1 = 10, L_2, ..., L_{10} = 1$. This is because in the former problem the forecaster can select the arm at period 10 based on the reward information in periods $1, ..., 9$ while in the latter one there is no information at the first period and no way to avoid 10 round losses. On the other hand, the size of the lock-up periods is also of great interest. We use parentheses to denote size-sorted periods: "(1)" indicates $\arg\max_n L_n$ and "(2)" indicates the second largest, etc. $L_{(1)}$ is also denoted as $L_{\max}$.

4.2.3 Related work

Multi-armed bandit problems have been extensively studied in the area of the machine learning and the operations research due to their simplicity and wide applications. The stochastic multi-armed bandit problem, in which the rewards of arms are drawn from some distributions, has attracted much attention. UCB1 [Auer et al., 2002a] is an efficient algorithm and is widely used.

Interesting problems that pose restrictions on forecaster's selection have been investigated. The bandit problem with switching costs is extensively studied. In this problem, the switching of arms generates a certain amount of loss, and the forecaster is motivated to stay with the current decision. For further details, see [Jun, 2004, Mahajan and Teneketzis, 2008, Guha and Munagala, 2009]. Note that, the switching cost is a soft constraint in the sense that the algorithm can switch the arm by paying some cost. However, in the examples explained in Section 4.1, changing an arm is difficult no matter how much cost the algorithm pays. In other words, the constraint is rather hard: the lock-up restriction in this thesis is motivated by this fact.

Motivated by experimental design settings, many papers considered a two-phase bandit problem, where the forecaster can select the arms freely in the experimental phase but must commit to a single arm in the terminal phase. Colton [1963] is the one of the oldest papers on this problem, and an extensive list of the studies on this problem can be found in Berry and Fristedt [1985]. Committing bandit [Bui et al., 2011] is one of the modern versions of the two-phase bandit problem. Three settings were investigated in Bui et al. [2011] for the length of the experimental phase. For two of the three settings, the forecaster can extend the experimental phase with a certain amount of cost, and the main

result of the paper is the algorithms for finding the optimal time to end the experimental phase. For the third setting[*3], the experimental phase has a fixed length N_e, and they showed optimal algorithm up to a constant factor. The third setting is equal to lock-up bandit with periods $L_1, ..., L_{N_e} = 1$ and $L_{N_e+1} = T - N_e + 1$. Our lock-up bandit can be considered as a generalization of these two-phase bandit problems: it does not separate the experimentation phase and commitment phase. Algorithms for lock-up bandit problem is applicable to any sizes lock-up periods restriction.

Lock-up bandit is also related to the best arm identification problem with fixed budget [aud]. In the best arm identification problem, the task of the forecaster is to find the best arm (optimal arm) among K arms. There is a fixed test period, and immediately after the end of the test period the forecaster outputs a "recommendation" arm he believes is the best. In the test period, the forecaster can select the arm at each round freely and receives the rewards. The test period has a fixed length d and the forecaster is evaluated based on the probability that the recommendation arm he selects corresponds to the real best arm. This setting is equal to the lock-up bandit with $L_1, ..., L_d = 1$ and $L_{d+1} \to \infty$: When the last period is sufficiently large, the regret in the test period is negligible.

4.3 Conversion from Stochastic Bandit Algorithms

There have been extensive study on the stochastic bandit and many algorithms have been proposed. Stochastic bandit algorithms assume that, at every round the forecaster can choose an arm freely. However, once the choice is restricted, it is not clear how to determine the next arm. In this section, we discuss the simple conversion from stochastic bandit algorithms into lock-up bandit algorithms. We also show that the converted UCB's regret is $O(\log T + L_{\max})$.

Proposition 4.3.1. (Conversion from stochastic bandit algorithms to lock-up bandit algorithms)

Let $\mathcal{A}$ be a stochastic bandit algorithm. In the following, we define an algorithm $\mathcal{A}'$ for the lock-up bandit that uses $\mathcal{A}$ as an internal algorithm. On one hand, if each round is the start of the lock-up period, $\mathcal{A}'$ invokes $\mathcal{A}$ and receives an arm. Then uses the arm as $\mathcal{A}$'s selection. On the other hand, if the round is not the start of the lock-up period, $\mathcal{A}'$ selects the same arm as the last round. In this case, invoke $\mathcal{A}$ and discard its selection. After receiving a reward, $\mathcal{A}'$ feeds $\mathcal{A}$ the selection and reward tuple $(I_t, \widehat{X}_{I(t)}(t))$. $\mathcal{A}$ learns from the reward tuple as if it were selected by itself.

It is true that the conversion above is not promised to be applicable for all algorithms in stochastic bandit[*4]. However, most algorithms, including confidence bound based al-

[*3] The authors called it a "hard experimentation deadline setting."

[*4] For example, for algorithms that maintain lists and select the next arms from the lists, the conversion

gorithms (UCB1, UCB-Tuned [Auer et al., 2002a], UCB-E [aud], UCB-V [Audibert et al., 2009], MOSS [Audibert and Bubeck, 2009], KL-UCB [Garivier and Cappé, 2011], etc.), and ϵ_n-greedy can be converted into lock-up bandit algorithms with the above procedure.

We denote converted algorithms using primes. For example, UCB1 and ϵ_n-greedy converted are UCB1′ and ϵ_n-greedy′. Remember that our main concern is the regret in lock-up bandit.

Theorem 13. (Regret upper bound of UCB1′) *The regret of UCB1′ in lock-up bandit is upper bounded as*

$$\mathbb{E}[\mathrm{Reg}(L_1, ..., L_N)] \leq \sum_{i \neq i^*} \left\{ \frac{8 \log T}{\Delta_i} + L_{\max} \Delta_i \left(1 + \frac{\pi^2}{3} \right) \right\}.$$

Proof sketch: the proof is the extension of the one in Auer et al. [2002a] to the lock-up bandit. The base theorem relies on the fact that the probability of suboptimal arm i played after $T_i(t) \geq \lceil (8 \log T)/\Delta_i^2 \rceil$ is sufficiently low and its sum is loosely bounded by $\pi^2/3$. In lock-up bandit, there are two main changes,

(1) $(8 \log T)/\Delta_i^2$ is replaced with $(8 \log T)/\Delta_i^2 + (L_{\max} - 1)$. The number of arms selected before $T_i(t) \geq (8 \log T)/\Delta_i^2$ is upper bounded by this quantity.
(2) $\pi^2/3$ is multiplied by $L_{\max}$.

The full proof is presented in Section 4.7.

Theorem 13 indicates that, the regret of UCB1′ is bounded by $O(\log T + L_{\max})$ for any list of lock-up periods $L_1, ..., L_N, \sum_n L_n = T$. When $L_{\max}$ is small compared with $\log T$, UCB1′ achieves $O(\log T)$ regret. Since the optimal regret bound in the lock-up bandit is logarithmic (c.f. Remark 12), the bound is optimal up to a constant factor. However, when there are some periods that are bigger than the order of $\log T$, the regret in the periods matters. In the next section, we propose a meta-algorithm to reduce the regrets in large periods.

4.4 How to Reduce Regret in Large Periods

In this section, we propose BaR, a general meta-algorithm for reducing regrets in large periods.

4.4.1 Minimizing regret in large periods

In the lock-up bandit problem, an algorithm cannot change the arm during a lock-up period. If an algorithm selects a suboptimal arm i at the start of round n, the regret is increased by $\Delta_i L_n$. For this reason, we want to avoid choosing a suboptimal arm

above is not directly applicable.

at the start of large periods. The notion of simple regret introduced by Bubeck et al. [2009] describes the minimum possible regret in a specific round. They proposed a pure exploration bandit problem. In this problem, at each round the algorithm selects an arm and receives a reward. After receiving the reward, the algorithm outputs an additional arm: the recommendation arm. At a certain round the game ends, and the algorithm is evaluated based on the quality of the recommendation arm. The goal with the algorithm is to minimize the simple regret, or the one-time regret of the recommendation arm. In summary, the pure exploration bandit problem is the same framework as the stochastic bandit problem except for the existence of the recommendation arm and the goal. In terms of the exploration and the exploitation trade-off, recommendation arm is an exploitation-only arm. The simple regret describes the best possible accuracy of the recommendation arm. In contrast with the simple regret, the sum of rewards during the game is called a cumulative regret, which is the quantity to optimize in the stochastic bandit. We denote the cumulative regrets as $\mathrm{Reg}(T)$ and the simple regret as $\mathrm{reg}(T)$ (in the period-wise notation, $\mathrm{Reg}(L_1, ..., L_N)$ and $\mathrm{reg}(L_1, ..., L_N)$, respectively). Interestingly, there is a trade-off between the two regrets.

Theorem 14. (Cumulative regret and simple regret trade-off [Bubeck et al., 2009]) *For any bandit algorithm and any function $\xi = \xi(t)$, if there exists some constant C and the allocation algorithm satisfies*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq C\xi(T),$$

for all Bernoulli reward distributions $\{P_1, ..., P_K\}$, then the simple regret of any recommendation strategies based on the bandit algorithm has the following lower bound: there exists a constant D and

$$\mathbb{E}[\mathrm{reg}(T)] \geq \frac{\Delta}{2} \exp\left(-D\xi(T)\right).$$

An intuitive explanation of Theorem 14 is as follows: the minimum possible simple regret for a round is determined by the cumulative regret to that point. Bubeck et al. [2009] proposed three natural recommendation algorithms: Empirical Best Arm (EBA), Most Played Arm (MPA) and Empirical Distribution of Plays (EDP). We use EBA, which recommends the arm of the best empirical mean, throughout this chapter.

4.4.2 BaR meta-algorithm

Good algorithms of the multi-armed bandit problem balance exploration and exploitation and result in $O(\log T)$ expected cumulative regret. If there are lock-up periods, this balance is perturbed by $O(L_{\max})$. If the value of exploration becomes large (i.e., a sub-optimal arm is chosen at the start of the largest period ($L_{\max} \gg \log T$)), it is difficult to restore the optimal balance of exploration and exploitation. The main idea of the BaR meta-algorithm (Algorithm 7) is using the recommendation arms as its selection

Algorithm 7 BaR meta-algorithm

Require: K arms, $L_1, ..., L_N$, $N_r \in \mathbb{N}$, and base algorithm $\mathcal{A}$

for $n \in 1, ..., N$ **do**
  if $n \in \{(1), ..., (N_r)\}$ **then**
    use the recommendation arm ψ (EBA arm)
    select arm $I_n = \psi$
    receive reward X until the period ends. The reward information is discarded.
  else
    invokes $\mathcal{A}$ to query for the arm selection ϕ
    select arm $I_n = \phi$
    receive reward X and feed $\mathcal{A}$ with the reward tuple (I_n, X) until the period ends.
  end if
end for

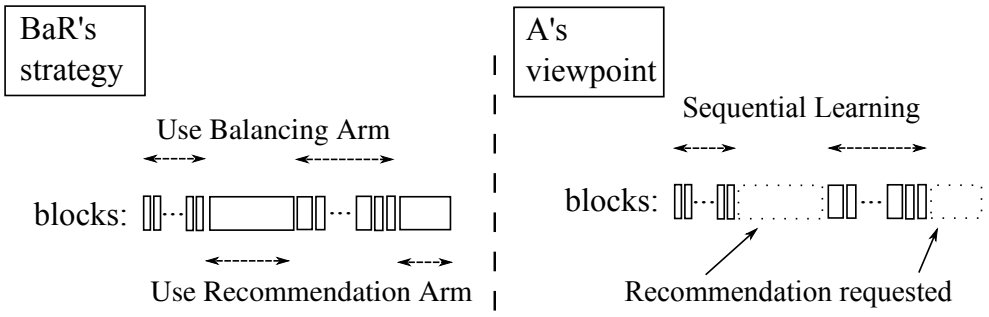


Fig. 4.2. BaR meta-algorithm. Two large periods are assigned to the recommendation set.

at large periods to avoid choosing suboptimal arms. This meta-algorithm uses a base lock-up bandit algorithm, which we denote as $\mathcal{A}$. Before the start of BaR, we decide the recommendation set $\{(1), ..., (N_r)\}$, or the top-N_r subset of the lock-up periods sorted by size. If each period is in the recommendation set, the algorithm queries $\mathcal{A}$ for the recommendation arm and uses it as the selection. $\mathcal{A}$ is not notified of the reward information. Otherwise, the algorithm works as a wrapper of $\mathcal{A}$. From the viewpoint of $\mathcal{A}$, it seems as if the periods in the recommendation set were banished (Figure 4.2). The regret of BaR can be derived from $\mathcal{A}$'s cumulative and simple regrets.

Remark 15. (Regret of [BaR, $\mathcal{A}$]) *If BaR is run with the recommendation set* $\{(1), ..., (N_r)\}$*, the regret is denoted by the base algorithm's cumulative and simple regret*

as,

$$\mathrm{Reg}(L_1, ..., L_N) =$$

$$\mathrm{Reg}_{base}(L_1, ..., L_N \setminus L_{(1)}, ..., L_{(N_r)}) + \sum_{n=1}^{N_r} L_{(n)} \mathrm{reg}_{base}(L_1, \ldots, L_{(n)-1} \setminus \{(1), ..., (N_r)\}),$$

where, the first term of RHS is the cumulative regret of $\mathcal{A}$ run in the environment where $\{(1), ..., (N_r)\}$ are removed. Also, in the second term of RHS, $L_1, \ldots, L_{(n)-1} \setminus \{(1), ..., (N_r)\}$ denotes the list of periods before the period (n) and not in $\{(1), ..., (N_r)\}$. For example, suppose $N = 100$ and the recommendation set $\{(1), ..., (N_r)\}$ is $\{(1), (2)\} = \{50, 100\}$. The cumulative regret is defined as the regret of the base algorithm run at the lock-up periods $1, ..., 49, 51, ..., 99$. The sum of simple regret is 50's simple regret after periods $1, ..., 49$ and period 100's simple regret after periods $1, ..., 49, 51, ..., 99$. The BaR meta-algorithm decomposes the regrets into the cumulative and the simple ones. The cumulative regret is dependent upon the maximum size of the periods (Theorem 13). By removing large periods, we can reduce the maximum size of the periods. Also, the recommendation is the best method for selecting the optimal arm. Therefore, it can minimize the regret generated from the simple regret part.

Our next concern is how to estimate the cumulative and simple regrets of base algorithms. In Section 4.3, we defined the uniform upper bound of a cumulative regret of UCB1′. However, we have not introduced any simple regret so far. In the next subsection, we describe UCB-E and discuss its regret.

4.4.3 UCB-E

UCB-E was introduced by aud as an explorative algorithm for stochastic bandit. It uses $\sqrt{a/T_i(t)}$ as the confidence bound. In the fixed horizon bandit game (i.e. T is known), the algorithm is flexible. When we set $a = 2\log T$, we obtain exactly the same the cumulative regret upper bound as UCB1 and can choose a large value to obtain a better simple regret bound. We convert UCB-E by using the Proposition procedure 4.3.1 to obtain UCB-E′.

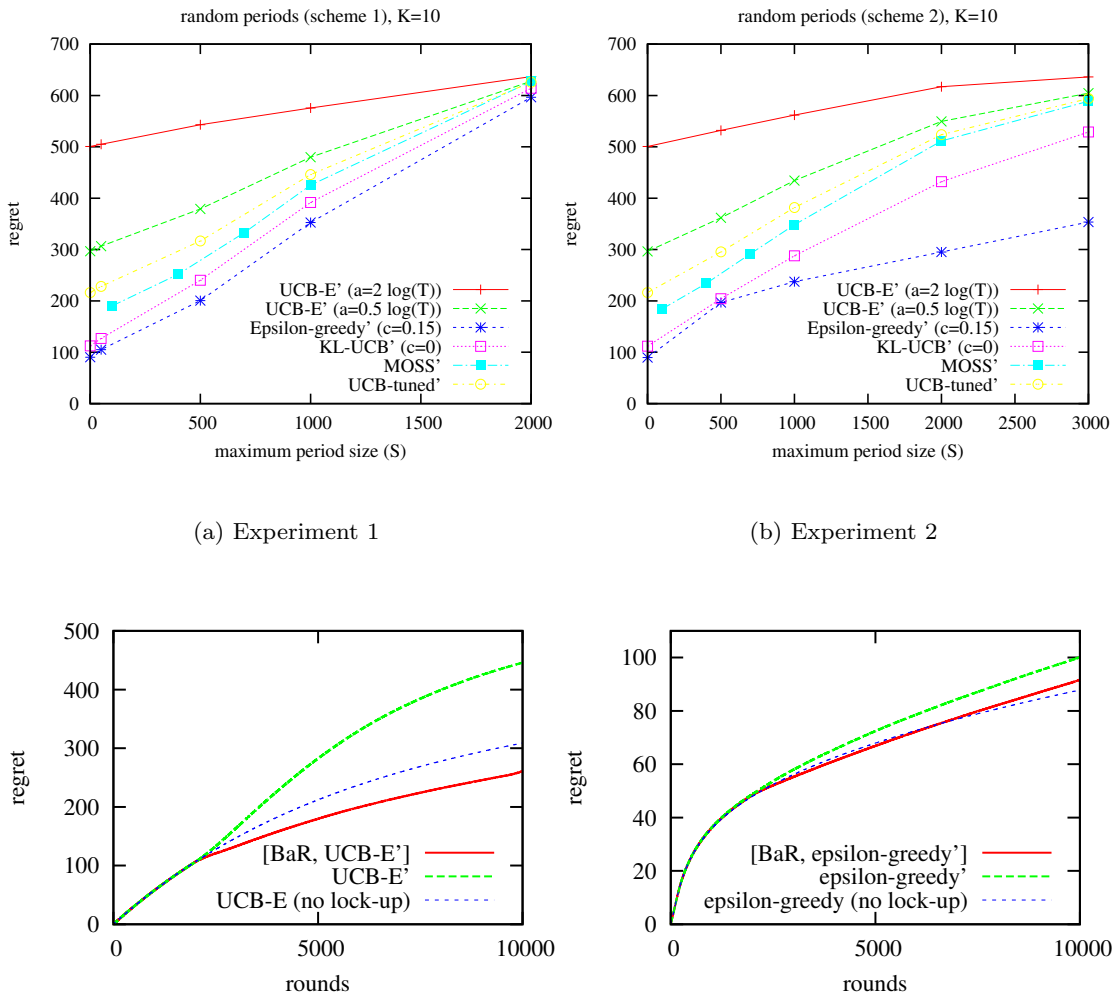
Theorem 16. (Regret upper bound of UCB-E′) *If UCB-E′ is run with parameter* $a \geq 2\log T$*, it satisfies*

$$\mathbb{E}[\mathrm{Reg}(L_1, ..., L_N)] \leq \sum_{i \neq i^*} \left\{ \frac{4a}{\Delta_i} + \Delta_i L_{\max} \left(1 + \frac{\pi^2}{3}\right) \right\}.$$

The proof is very similar to Theorem 13. The proof is presented in Section 4.7.

Theorem 17. (Simple regret upper bound of UCB-E′) *If UCB-E′ is run with parameter* $0 < a \leq \frac{25}{36} \frac{T - K L_{\max}}{H_1}$ *then it satisfies*

$$\mathbb{E}[\mathrm{reg}(L_1, ..., L_N)] \leq 2TK \exp\left(-\frac{2a}{25}\right),$$



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

(d) Experiment 4

Fig. 4.3. Experimental results. Experiments 1 and 2 show regret as function of maximum period size. Experiments 3 and 4 show regret before/after application of BaR.

where, $H_1 = \sum_{i \neq i^*} 1/\Delta_i^2 + 1/\Delta^2$.

Proof Sketch: the proof relies on the fact that the empirical mean rarely deviates from the thin confidence bound $(1/5)\sqrt{a/T_i(t)}$. It holds in all $a \leq 25(T - KL_{\max})/(36H_1)$, even in the existence of lock-up periods. The full proof is presented in Section 4.7.

4.5 Experiments

We conducted two sets of experiments to support the theoretical results in the previous two sections.

(1) In Section 4.3, we proposed a simple conversion from a stochastic bandit algorithm

to a lock-up bandit algorithm. The converted algorithm's regret is linearly dependent upon the maximum period size. In the first set of experiments (Experiments 1 and 2), we studied the dependency between the maximum period size and a regret.

(2) In Section 4.4, we proposed the BaR meta-algorithm, which reduces the regret in large periods. In the second set of experiments (Experiments 3 and 4), we conducted a before/after analysis of BaR.

4.5.1 Experimental settings

All experiments involved ten-armed lock-up bandits with $T = 10000$. The rewards of arms were Bernoulli distributions with means
$(\mu_1, ..., \mu_{10}) = (0.1, 0.05, 0.05, 0.05, 0.02, 0.02, 0.02, 0.01, 0.01, 0.01)$.

Settings of Experiments 1 and 2

The algorithms we used were UCB-E$'$ [aud] with parameter $a = 2 \log T$, $a = 1/2 \log T$, ϵ_n-greedy$'$ [Auer et al., 2002a] with parameters $(c, d) = (0.15, 0.1)$, MOSS$'$ [Audibert and Bubeck, 2009], KL-UCB$'$ [Garivier and Cappé, 2011] with parameter $c = 0$, and UCB-Tuned$'$ [Auer et al., 2002a]. We do not intend to argue which algorithm is better[*5].

We showed the regret as a function of maximum period size S ([*6]). In all experiments, for each value of S we show an averaged regret over 10,000 different runs. For each run, the lock-up periods in the experiments were randomly generated as follows. Until the total number of rounds reached T (i.e., $\sum_n L_n < T$), we appended a new period of size $\{1, ..., S\}$ with the same probability (Experiments 1) or the probability proportional to the inverse of size (Experiments 2). The last period was decreased to satisfy $\sum_n L_n = T$.

Settings of Experiments 3 and 4

In the second set of experiments (Experiments 3 and 4), we showed the regret as a function of rounds. The algorithms we used were UCB-E$'$ with parameter $a = 1/2 \log T$ (Experiment 3) and ϵ_n-greedy$'$ with parameters $(c, d) = (0.15, 0.1)$ (Experiment 4). In both experiments, the regrets were averaged over 10,000 different runs. In each run, the periods were generated as follows. For the first 2,000 rounds there were no lock-up periods (i.e., $L_1, ..., L_{2000} = 1$). From rounds 2,001 to 10,000, the periods were generated randomly as follows. Until the sum of the periods reached 10,000, we appended a new period of size $\{1, ..., 1000\}$ with the probability proportional to the inverse of size. We compared the base algorithm (UCB-E$'$ and ϵ_n-greedy$'$) before and after application of BaR. We also show the regret of the base algorithm run with no lock-up period (= the stochastic

[*5] The parameters in ϵ_n-greedy$'$ were chosen to be empirically good (c.f. Section 4 in [Auer et al., 2002a]). Therefore, it was no surprise ϵ_n-greedy$'$ performed better than UCB-E$'$.

[*6] S is the maximum period size to be possibly generated. $L_{\max}$, the maximum period size to be actually generated, is smaller than or equal to S.

bandit), which is much easier than the lock-up bandit. As for the recommendation set, we used all periods larger than 400.

4.5.2 Experimental results and discussions

Results of Experiments 1 and 2

Figure 4.3 is the results of the experiments. In Experiments 1 and 2, we observed linear relation between the maximum period size and the regret for all algorithms. Note that, between Experiments 1 and 2, the number of large periods differed greatly. In Experiment 2, large periods had a small probability (inverse to its size) to be generated compared with Experiment 1; however, the results of Experiments 1 and 2 look very much alike. This fact supports that the regret in lock-up bandit is dependent upon the size of the maximum periods.

Results of Experiments 3 and 4

Experiments 3 and 4 showed the effect of BaR. In both experiments, using BaR makes the regret significantly smaller. In Experiments 3, the results of [BaR, UCB-E$'$] were even better than those of the base algorithm in the standard bandit game. This is surprising because the bandit problem with lock-up periods is much more difficult than the standard bandit problem. This can be explained as follows. The regret of UCB-E$'$ is higher than that of ϵ_n-greedy$'$. This means that UCB-E$'$ does more exploration than it should and there is some room for exploitation.number of recommendation periods In the Experiments 4, the regret of [BaR, ϵ_n-greedy$'$] was higher than that of no lock-ups, which is natural. This results are not specific to UCB-E and ϵ_n-greedy. We also conducted experiments with many state-of-the-art algorithms (KL-UCB [Garivier and Cappé, 2011], MOSS [Audibert and Bubeck, 2009] and UCB-Tuned [Auer et al., 2002a]) and obtained similar results.

Discussions

The use of the BaR meta-algorithm effectively reduces regret for the following reason. When T is large, the ratio of exploration to exploitation is small (i.e. $O(\log T/T) \to 0$). Therefore, if the forecaster does more exploration than it should do, restoring the optimal balance is virtually impossible. Conversely, if it does less exploration is smaller than it should, restoring the optimal balance is relatively easy. This is why BaR, which increases exploitation during the large lock-up periods, works well.

4.6 Conclusion and Future Works

We proposed and studied a bandit problem with a lock-up period restriction, which is expected to model the practical scenarios that naturally arise when we apply stochastic bandit to real problems. We studied how the exploration and exploitation balance is

perturbed by lock-up restrictions and proposed methods to recover the balance. For further understanding of related problems, better bounds for the simple regret is of great interest. Contrary to the cumulative regret, the simple regret is less known. In our theory, the simple regret is important and finer bound preferred.

4.7 Proofs

In this section, we prove the theorems in this chapter. The overall goal with the proofs is to show that the existing bounds in stochastic bandit also holds even in the existence of lock-up periods.

4.7.1 Array-UCB

The proofs of the cumulative regrets in UCB1 and UCB-E rely on the same bound. To avoid redundancy, we define Array-UCB, the generalization of UCB1 and UCB-E.

Definition 1. (Array-UCB) *Let $\widehat{\mu}_{i,s}$ be the empirical mean reward of arm i with s samples. Array-UCB with a real-valued function $a = a(t)$ is defined as the confidence bound based algorithm with the following index of each arm:*

$$\widehat{\mu}_{i,s} + \sqrt{\frac{a(t)}{s}}. \tag{4.1}$$

At each round t, the forecaster selects the arm of the maximum index

When $a(t) = 2\log t$, Array-UCB is equal to UCB1. When $a(t) = a$ (constant), Array-UCB is equal to UCB-E.

4.7.2 Proofs of Theorem 13 and Theorem 16

In this subsection, we prove Theorems 13 and 16, the uniform cumulative regret of UCB1′ and UCB-E′. We convert Array-UCB to an algorithm for lock-up bandit with the procedure of Proposition 4.3.1. We call the converted algorithm Array-UCB′.

Theorem 18. *For Array-UCB′ with $a(t) \geq 2\log t$, the cumulative regret in the lock-up bandit problem is upper bounded as*

$$\mathbb{E}[\mathrm{Reg}(L_1, ..., L_N)] \leq \sum_{i \neq i^*} \left\{ \frac{4a_{\max}}{\Delta_i} + \Delta_i L_{\max} \left(1 + \frac{\pi}{3}\right) \right\},$$

where $a_{\max} = \max_{t \in \{1,...,T\}} a(t)$.

Theorems 13 and 16 are directly derived as the specialization of Theorem 18 with $a(t) = 2\log t$ and $a(t) = a \geq 2\log T$.

Proof of Theorem 18. The proof is based on Proof 1 in Auer et al. [2002a]. The base proof is on UCB1 in the stochastic bandit. We extend this proof in two respects. First, the

UCB1 index is generalized to the Array-UCB index (Equation (4.1)). Second, we take lock-up periods into consideration.

We upper bound $T_i(T)$, the number of rounds suboptimal arm i is pulled in T rounds. Let $c_{t,s} = \sqrt{a(t)/s}$. Remember (s_n, f_n) tuple means the start and end round of the period n. We use both the period-wise notation with symbol n and round-wise notation with symbol t (c.f. Section 4.2.1).

$$\begin{aligned} T_i(L_1, ..., L_N) &= \sum_{n=1}^{N} L_n \mathbf{1}\{I_n = i\} \\ &= (l + L_{\max} - 1) + \sum_{n=K+1}^{N} L_n \mathbf{1}\{I_n = i, T_i(s_n - 1) \geq l\}, \end{aligned} \tag{4.2}$$

where the transformation at (4.2) comes from the fact that T_i is at most $l + L_{\max} - 1$ at the first period after T_i exceeds or equals l. The condition i is selected at $n \geq 2$ is transformed as follows.

$$\begin{aligned} \mathbf{1}\{I_n = i\} &\leq \mathbf{1}\{\widehat{\mu}_{i^*, T_{i^*}(s_n - 1)} + c_{s_n - 1, T_{i^*}(s_n - 1)} \leq \widehat{\mu}_{i, T_i(s_n - 1)} + c_{s_n - 1, T_i(s_n - 1)}\} \\ &\leq \mathbf{1}\{\min_{0 < t_1 < s_n} \widehat{\mu}_{i^*, t_1} + c_{s_n - 1, t_1} \leq \max_{l < t_2 < s_n} \widehat{\mu}_{i, t_2} + c_{s_n - 1, t_2}\} \\ &\leq \sum_{t_1 = 1}^{s_n - 1} \sum_{t_2 = 1}^{s_n - 1} \mathbf{1}\{\widehat{\mu}_{i^*, t_1} + c_{s_n - 1, t_1} \leq \widehat{\mu}_{i, t_2} + c_{s_n - 1, t_2}\}. \end{aligned} \tag{4.3}$$

The condition $\widehat{\mu}_{i^*, t_1} + c_{s_n - 1, t_1} \leq \widehat{\mu}_{i, t_2} + c_{s_n - 1, t_2}$ in (4.3) implies that at least one of the following three conditions must hold.

$$\widehat{\mu}_{i^*, t_1} \leq \mu^* - \sqrt{\frac{a(s_n - 1)}{t_1}}, \tag{4.4}$$

$$\widehat{\mu}_{i, t_2} \geq \mu_i + \sqrt{\frac{a(s_n - 1)}{t_2}}, \tag{4.5}$$

$$\mu^* < \mu_i + 2\sqrt{\frac{a(s_n - 1)}{t_2}}. \tag{4.6}$$

Now, we bound the probabilities of inequalities (4.4), (4.5), and (4.6). First, (4.6) never occurs when $t_2 \geq \lceil \frac{4a_{\max}}{\Delta_i^2} \rceil$. The probability of (4.4) is upper bounded as

$$\begin{aligned} \mathbb{P}[(4.4)\,\text{is true}] &= \mathbb{P}\left[\widehat{\mu}_{i^*}(t_1) \leq \mu^* - \sqrt{\frac{a(s_n - 1)}{t_1}}\right] \\ &\leq \mathbb{P}\left[\widehat{\mu}_{i^*, t_1} \leq \mu^* - \sqrt{\frac{2\log(s_n - 1)}{t_1}}\right] \qquad (4.7) \\ &\leq \exp(-4\log(s_n - 1)) \leq (s_n - 1)^{-4}, \qquad (4.8) \end{aligned}$$

where we use the assumption $a(t) > 2\log t$ at (4.7) and the Hoeffding's inequality at (4.8). By using the same arguments, we obtain the same bound for (4.5). By using inequalities (4.2), (4.3) and (4.8), we obtain

$$\begin{aligned}
\mathbb{E}[T_i(N)] &\leq \left(\left\lceil \frac{4a_{\max}}{\Delta_i^2} \right\rceil + L_{\max} - 1\right) \\
&\quad + \sum_{t_1=1}^{s_n-1} \sum_{t_2=1}^{s_n-1} \mathbf{1}\left\{ \widehat{\mu}_{i^*,t_1} + c_{s_n-1,t_1} \leq \widehat{\mu}_{i,t_2} + c_{s_n-1,t_2}, t_2 \geq \left(\left\lceil \frac{4a_{\max}}{\Delta_i^2} \right\rceil + L_{\max} - 1\right)\right\} \\
&\leq \left(\left\lceil \frac{4a_{\max}}{\Delta_i^2} \right\rceil + L_{\max} - 1\right) \\
&\quad + \sum_{n=K+1}^{N} \sum_{t_1=1}^{s_n-1} \sum_{t_2=1}^{s_n-1} L_s \Big\{ \mathbb{P}[((4.4) \text{ is true})] + \mathbb{P}[((4.5) \text{ is true})] \Big\} \\
&\leq \left(\frac{4a_{\max}}{\Delta_i^2} + L_{\max}\right) + L_{\max} \sum_{n=K+1}^{N} \sum_{t_1=1}^{s_n-1} \sum_{t_2=1}^{s_n-1} (2(s_n-1)^{-4}) \\
&\leq \left(\frac{4a_{\max}}{\Delta_i^2} + L_{\max}\right) + L_{\max} \sum_{t=1}^{\infty} (2t^{-2}) \\
&\leq \left(\frac{4a_{\max}}{\Delta_i^2} + L_{\max}\right) + L_{\max} \cdot \frac{\pi^2}{3} = \frac{4a_{\max}}{\Delta_i^2} + L_{\max}\left(1 + \frac{\pi^2}{3}\right).
\end{aligned}$$

□

4.7.3 Proof of Theorem 17

Proof of Theorem 17. We extend Theorem 1 in aud to lock-up bandit. Consider an event

$$\xi = \left\{ \forall i \in \{1, ..., K\}, t \in \{1, ..., T\}, |\widehat{\mu}_{i,t} - \mu_i| < \frac{1}{5}\sqrt{\frac{a}{t}} \right\}.$$

By using the Hoeffding's inequality for each event and the union bound over rounds and arms, we have $\mathbb{P}(\xi) \geq 1 - 2TK\exp\left(-\frac{2a}{25}\right)$. Indeed, the event is the sufficient condition for that the empirically optimal arm corresponds to the truly optimal arm, as shown in the following argument. Assume that ξ holds. It is enough to prove that

$$\frac{1}{5}\sqrt{\frac{a}{T_i(T)}} \leq \frac{\Delta_i}{2}, \forall i \in \{1, ..., K\},$$

or equivalently

$$T_i(T) \geq \frac{4}{25}\frac{a}{\Delta_i^2}.$$

First, we prove the upper bound of the number of the suboptimal arms pulled, namely

$$T_i(t) \leq \frac{36}{25}\frac{a}{\Delta_i^2} + L_{\max}, \forall i \neq i^*. \tag{4.9}$$

Since the algorithm can select an arm only at the start of each lock-up period, we use induction based on each period. Namely, we show (4.9) is true at the end of any periods.

Remember, we denote the start and end of the lock-up period n as (s_n, f_n). We also denote the UCB-E index as $B_{i,s} = \widehat{\mu}_{i,s} + \sqrt{a/s}$. Obviously the inequality holds when $n = 1$. We now assume the inequality is true at the end of period $n-1$. If $I_n \neq i$, $T_i(f_n) = T_i(f_{n-1})$ and the inequality still holds at the end of period n. On the other hand, if $I_n = i$ then it means $B_{i,T_i(s_n-1)} \geq B_{i^*,T_{i^*}(s_n-1)}$. Since event ξ holds, we have $B_{i^*,T_{i^*}(s_n-1)} \geq \mu^*$ and $B_{i,T_i(s_n-1)} \leq \mu_i + \frac{6}{5}\sqrt{\frac{a}{T_i(s_n-1)}}$. Summing up these conditions, we obtain $\frac{6}{5}\sqrt{\frac{a}{T_i(s_n-1)}} \geq \Delta_i$. The arm i is chosen during the lock-up period n. Since $f_n - (s_n - 1) = L_n \leq L_{\max}$, (4.9) still holds.

Next, we prove the lower bound of suboptimal arms selected

$$T_i(t) \geq \frac{4}{25} \min\left(\frac{a}{\Delta_i^2}, \frac{25}{36}(T_{i^*}(t) - L_{\max})\right), \forall i \neq i^*. \tag{4.10}$$

We also use induction based on each lock-up period. We assume (4.10) holds at the end of period $n-1$. Then, at period n , if $B_{i,T_i(s_n-1)} > B_{i^*,T_{i^*}(s_n-1)}$, then T_{i^*} does not increase, so it still holds. On the other hand, in the case of $B_{i,T_i(s_n-1)} \leq B_{i^*,T_{i^*}(s_n-1)}$, T_{i^*} might increase. Since we are assuming event ξ,

$$\mu^* + \frac{6}{5}\sqrt{\frac{a}{T_{i^*}(s_n-1)}} \geq B_{i^*,T_{i^*}(s_n-1)} \geq B_{i,T_i(s_n-1)} \geq \mu_i + \frac{4}{5}\sqrt{\frac{a}{T_i(s_n-1)}},$$

which gives

$$T_i(s_n-1) \geq \frac{16}{25}\frac{a}{\left(\Delta_i + \frac{6}{5}\sqrt{\frac{a}{T_{i^*}(s_n-1)}}\right)^2}.$$

By using $u + v \leq 2\max(u, v)$, $T_i(f_n) = T_i(s_n - 1)$, and $T_{i^*}(f_n) \geq T_{i^*}(s_n - 1) + L_{\max}$, (4.10) holds at the end of period n. From (4.10), we only have to show that, for all $i \neq i^*$

$$\frac{25}{36}(T_{i^*}(T) - L_{\max}) \geq \frac{a}{\Delta_i^2}.$$

By using (4.9), we obtain

$$T_{i^*}(T) - L_{\max} = T - L_{\max} - \sum_{i \neq i^*} T_i(T) \geq T - KL_{\max} - \frac{36}{25}a\sum_{i \neq i^*}\Delta_i^{-2} \geq \frac{36}{25}a\Delta^{-2},$$

where, we use the assumption of the theorem, $\frac{36}{25}H_1 a \geq T - KL_{\max}$ in the last inequality. □

Chapter 5

Asymptotically Optimal Exploration and Exploitation in Multiple-play Multi-armed Bandit Problem

In this chapter, we discuss a multiple-play multi-armed bandit problem in which several arms are selected at each round. We propose multiple-play Thompson sampling (MP-TS) algorithm, an extension of Thompson sampling (TS) to the multiple-play MAB problem, and analyze its regret. We prove that MP-TS for binary rewards has a regret upper bound that matches the asymptotic regret lower bound provided by Anantharam et al. [1987]. A set of computer simulations was also conducted, which compared MP-TS with state-of-the-art algorithms. We also propose a modification of MP-TS, which is shown to have better empirical performance[*1]. The notation in this chapter is summarized in Table 5.1.

5.1 Motivation

Up to the previous chapter, we have specifically dealt with the bandit problem in which an arm is selected and drawn at each round. Let us call this problem single-play (SP) bandit. While the SP bandit problem is indisputably important as a canonical problem, in many practical situations multiple entities corresponding to arms are selected at each round. We call the bandit problem in which several arms can be selected multiple-play (MP) bandit. Examples of the situations that can be modeled as an MP bandit problem include the followings.

- **Example 1 (placement of online advertisements):** a website has several slots where advertisements can be placed. Based on each user's query, there is a set of candidates of relevant advertisements from which websites can select to display.

[*1] The contents of this chapter were published in Komiyama et al. [2015b].

Table. 5.1. Notation used in Chapter 5.

$\mathbf{1}\{A\}$	:=	1 if A is true and 0 otherwise.
K	:=	Number of the arms.
$[K]$	:=	$\{1, 2, \ldots, K\}$.
L	:=	Number of the selections at each round.
T	:=	Number of the rounds.
$I(t)$	:=	Set of the arms that is selected in round t.
$\widehat{X}_i(t)$	:=	Reward of arm i at round t.
μ_i	:=	Expected reward of arm i. In this chapter we assume $\mu_1 > \mu_2 > \cdots > \mu_K$.
ν	:=	$(\mu_{L-1} + \mu_L)/2$.
$\mu_L^{(-)}$	:=	$\mu_L - \delta$.
$\mu_i^{(+)}$	:=	$\mu_i + \delta$.
$\widehat{\mu}_i(t)$	:=	Empirical mean reward of arm i at round t.
$\widetilde{\mu}_i(t)$	:=	Posterior sample of arm i at round t.
$\widetilde{\mu}^*(t)$	:=	The L-th largest posterior sample among $\{\widetilde{\mu}_i(t)\}$.
$\widetilde{\mu}^{**}_{\setminus i,j}(t)$	:=	The $(L-1)$-th largest posterior sample at round t except for arms i and j.
$\Delta_{i,j}$	:=	$\mu_j - \mu_i$.
$N_i(t)$	:=	$\sum_{t'=1}^{t-1} \mathbf{1}\{i \in I(t')\}$.
$N_i^{\mathrm{suf}}(T)$	:=	$\log T / d(\mu_i^{(+)}, \mu_L^{(-)})$.
$d(p, q)$	:=	The KL divergence between Bernoulli distributions: $p \log(p/q) + (1-p) \log((1-p)/(1-q))$.

The effectiveness of advertisements varies: some advertisements are more appealing to the user than others. With the standard model in online advertising, it is assumed that each advertisement is associated with a click-through rate (CTR), which is the number of clicks per view. Since websites receive revenue from clicks on advertisements, it is natural to maximize it, which can be considered as an instance of an MP-MAB problem in which advertisements and clicks correspond to arms and rewards, respectively.

- **Example 2 (channel selection in cognitive radio networks [Huang et al., 2008]):** a cognitive radio is an adaptive scheme for allocating channels, such as wireless network spectrums. There are two kinds of users: primary and secondary. Unlike primary users, secondary users do not have primary access to a channel but can take advantage of the vacancies in primary access and opportunistically exploit instantaneous spectrum availability when primary users are idle. However, the availabilities of channels are not easily known. Usually, secondary users have access

to multiple channels. They can enhance their communication efficiency by adaptively estimating the availability statistics of the channels, which can be considered as an MP bandit problem in which channels and the permission of communication are arms and rewards, respectively.

There have been several studies on the MP bandit problem. Anantharam et al. [1987] derived an asymptotic lower bound on the regret for this problem and proposed an algorithm with a matching regret bound. Because their algorithm requires certain statistics that are difficult to compute, efficiently computable MP bandit algorithms have also been extensively studied. Chen et al. [2013] extended a UCB-based algorithm to a multiple-play case with combinatorial rewards and Gopalan et al. [2014] extended TS to a wide class of problems. Although both papers provide a logarithmic regret bound, the leading constant of these regret bounds do not match the lower bound. Therefore, it is unknown whether an asymptotically optimal regret bound for the MP bandit problem is achievable by using a computationally efficient algorithm. Note also that, there is recently another line of work called semi-bandits [Neu and Bartók, 2013, Wen et al., 2015] in which a subset of the arms are selected in each round.

The main difficulty in analyzing the MP bandit problem lies in the fact that the regret depends on the combinatorial structure of arm draws. More specifically, an algorithm with an asymptotically optimal bound on the number of draws of suboptimal arms does not always ensure the optimal regret unlike the SP bandit problem.

Contribution: our contributions are as follows.

- **TS-based algorithm for the MP bandit problem and its optimal regret bound:** the first and main contribution of this chapter is an extension of TS to the multiple-play case, which we call MP-TS. We prove that MP-TS for binary rewards has an asymptotically optimal regret bound.
- **Novel analysis technique:** to solve the difficulty in the combinatorial structure of the MP bandit problem, we show that the independence of posterior samples among arms in TS is a key property for suppressing the number of simultaneous draws of several suboptimal arms, and the use of this property eventually leads to the optimal regret bound.
- **Experimental comparison among MP bandit algorithms:** we compare MP-TS with other algorithms, and confirm its efficiency. We also propose an empirical improvement of MP-TS (IMP-TS) motivated by analyzes on the regret structure of the MP bandit problem. We confirm that IMP-TS improves the performance of MP-TS without increasing computational complexity.

5.2 Problem Setup

Let there be K arms. Each arm $i \in [K] := \{1, 2, \ldots, K\}$ is associated with a probability distribution $P_i = \text{Bernoulli}(\mu_i)$, $\mu_i \in (0, 1)$. At each round $t = 1, 2, \ldots, T$, the forecaster selects a set of $L < K$ arms $I(t)$, then receives the rewards of the selected arms. The reward $\widehat{X}_i(t)$ of each selected arm i is i.i.d. samples from P_i. Let $N_i(t)$ be the number of draws of arm i before round t (i.e., $N_i(t) = \sum_{t'=1}^{t-1} \mathbf{1}\{i \in I(t')\}$, where $\mathbf{1}\{\mathcal{A}\} = 1$ if event $\mathcal{A}$ holds and $= 0$ otherwise.), and $\widehat{\mu}_i(t)$ be the empirical mean of the rewards of arm i at the beginning of round t. The forecaster is interested in maximizing the sum of rewards over drawn arms. For simplicity, we assume that all arms have distinct expected rewards (i.e., $\mu_i \neq \mu_j$ for any $i \neq j$). We discuss the case in which $\mu_i = \mu_j$ for some i and j in Section 5.9. Without loss of generality, we assume $\mu_1 > \mu_2 > \mu_3 > \cdots > \mu_K$. Of course, algorithms do not exploit this ordering. We define optimal arms as top-L arms (i.e., arms $[L]$), and suboptimal arms as the others (i.e., arms $[K] \setminus [L]$). The regret, which is the expected loss of the forecaster, is defined as

$$\text{Reg}(T) = \sum_{t=1}^{T} \left(\sum_{i \in [L]} \mu_i - \sum_{i \in I(t)} \mu_i \right).$$

The expectation of regret $\mathbb{E}[\text{Reg}(T)]$ is used to measure the performance of an algorithm.

5.3 Regret Bounds

In this section we introduce the known lower bounds of the regret for the SP bandit and MP bandit problems and discuss the relation between them.

5.3.1 Regret bound for SP bandit problem

The SP bandit problem, which has been thoroughly studied in the fields of statistics and machine learning, is a special case of the MP bandit problem with $L = 1$. As we discussed in Section 2.3.2, the asymptotic regret lower bound in the SP bandit problem was given by Lai and Robbins [1985]. They proved that, for any strongly consistent algorithm (i.e., algorithms with subpolynomial regret for any set of arms), there exists a lower bound

$$\mathbb{E}[N_i(T+1)] \geq \left(\frac{1 - o(1)}{d(\mu_i, \mu_1)} \right) \log T, \tag{5.1}$$

where $d(p, q) = p \log(p/q) + (1-p) \log((1-p)/(1-q))$ is the KL divergence between two Bernoulli distributions with expectation p and q. Note that when arm i is drawn, the

regret increases by $\Delta_{i,1}$ and the regret is written as

$$\mathbb{E}[\mathrm{Reg}(T)] = \sum_{i \neq 1} N_i(T+1)\Delta_{i,1}, \tag{5.2}$$

where $\Delta_{i,j} = \mu_j - \mu_i$. Therefore, inequality (5.1) directly leads to the regret lower bound

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \sum_{i \neq 1} \left(\frac{(1 - o(1))\Delta_{i,1}}{d(\mu_i, \mu_1)} \right) \log T. \tag{5.3}$$

One may think that applying the techniques of the SP bandit problem would directly yield an optimal bound for a more general MP bandit problem. However, this is not the case. In short, the difficulty in analyzing the regret on the MP bandit problem arises from the fact that the optimal bound on the number of suboptimal arm draws does not directly lead to the optimal regret. From this point forward, we focus on the MP bandit problem in which L is not restricted to one.

5.3.2 Extension to MP bandit problem

The asymptotic regret lower bound in the MP bandit problem, which is the generalization of inequality (5.3), was provided by Anantharam et al. [1987]. They first proved that, for any strongly consistent algorithm and suboptimal arm i, the number of arm i draws is lower-bounded as

$$\mathbb{E}[N_i(T+1)] \geq \left(\frac{1 - o(1)}{d(\mu_i, \mu_L)} \right) \log T. \tag{5.4}$$

Unlike in the SP bandit problem, the regret in the MP bandit problem is not uniquely determined by the number of suboptimal arm draws. As illustrated in Figure 5.1, the regret is dependent on the combinatorial structure of arm draws.

Recall that a regret increase at each round is the gap of expected rewards between the optimal arms and that of the selected arms. When a suboptimal arm is selected, one optimal arm is excluded from $I(t)$ instead of the suboptimal arm. Let the selected suboptimal arm and excluded optimal arm be i and j, respectively. Then, we lose expected reward $\mu_j - \mu_i$. Namely, the loss in the expected reward at each round is given by

$$\begin{aligned} \sum_{j \in [L]} \mu_j - \sum_{i \in I(t)} \mu_i &= \sum_{j \in [L] \setminus I(t)} \mu_j - \sum_{i \in I(t) \setminus [L]} \mu_i \\ &\geq \sum_{i \in I(t) \setminus [L]} (\mu_L - \mu_i), \end{aligned} \tag{5.5}$$

where we used the fact $\mu_j \geq \mu_L$ for any optimal arm j. From this relation, the regret is

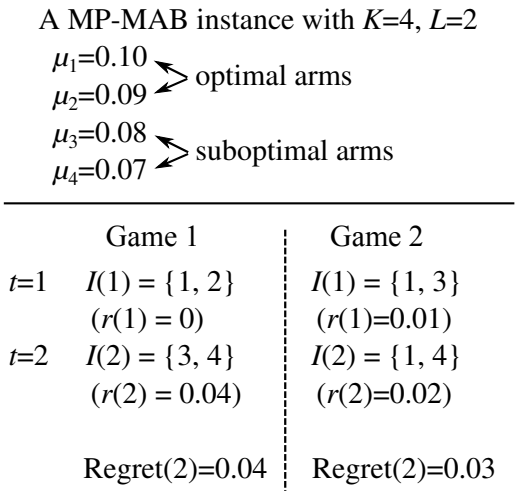


Fig. 5.1. Two bandit games with the same set of arms. $r(t)$ is defined as the increase in the regret at round t. In both games 1 and 2, we have the same number of suboptimal arm draws ($N_3(2) = N_4(2) = 1$). However, the regret in games 1 and 2 are different.

expressed as

$$\begin{aligned}\mathrm{Reg}(T) &\geq \sum_{t=1}^{T} \sum_{i \in I(t) \setminus [L]} (\mu_L - \mu_i) \\ &= \sum_{i \in [K] \setminus [L]} (\mu_L - \mu_i) N_i(T+1) \end{aligned} \tag{5.6}$$

which, combined with (5.4), leads to the regret lower bound by Anantharam et al. [1987] that any strongly consistent algorithm satisfies

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \sum_{i \in [K] \setminus [L]} \frac{(1 - o(1))\Delta_{i,L}}{d(\mu_i, \mu_L)} \log T. \tag{5.7}$$

5.3.3 Necessary condition for an optimal algorithm

In Sections 5.3.1 and 5.3.2, we saw that the derivations of the regret bounds are analogous between the SP bandit and MP bandit problems. However, there is a difference in the relation between the regret and $N_i(T)$, the number of draws of suboptimal arms, is given as equation (5.2) in the SP bandit problem, whereas it is given as inequality (5.6) in the MP bandit problem. This means that, an algorithm achieving the asymptotic lower bound (5.4) on $N_i(T)$ does not always achieve the asymptotic regret bound (5.7).

When suboptimal arm i is selected, one of the optimal arms is pushed out instead of arm i, and the regret increases by the difference between the expected rewards of these two arms. The best scenario is that, arm L, which is the optimal arm with the smallest expected reward, is almost always the arm pushed out instead of a suboptimal arm. For

Algorithm 8 Multiple-play Thompson sampling (MP-TS) for binary rewards

Input: # of arms K, # of selection L
for $i = 1, 2, \ldots, K$ **do**
  $A_i, B_i = 1, 1$
end for
$t \leftarrow 1$.
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, K$ **do**
    $\widetilde{\mu}_i(t) \sim \mathrm{Beta}(A_i, B_i)$
  end for
  $I(t) =$ top-L arms ranked by $\widetilde{\mu}_i(t)$.
  for $i \in I(t)$ **do**
    if $\widehat{X}_i(t) = 1$ **then**
      $A_i \leftarrow A_i + 1$
    else
      $B_i \leftarrow B_i + 1$
    end if
  end for
end for

this scenario to occur, it is necessary to ensure that at most one suboptimal arm is drawn for almost all rounds because, if two suboptimal arms are selected, at least one arm in $[L-1]$ is pushed out.

In the next section, we propose an extension of TS to the MP bandit problem, and explain that it has a crucial property for suppressing this simultaneous draw of two suboptimal arms.

Remark: Corollary 1 of Gopalan et al. [2014] shows the achievability of the bound in the RHS of (5.4) on the number of draws of suboptimal arms. Whereas this does not lead to an asymptotically optimal regret bound as discussed above, they originally derived in Theorem 1 an $O(\log T)$ bound on the number of each suboptimal action (that is, each combination of arms including suboptimal ones) for a more general setting of MP bandit. Thus, we can directly use this bound to derive a better regret bound. However, to show the optimality in the sense of regret it is necessary to prove that there are at most $o(\log T)$ rounds such that an arm in $[L-1]$ is pushed out. Therefore, it still requires further discussion to derive the optimal regret bound of TS. Note also that the regret bound by Gopalan et al. [2014] is restricted to the case that the prior has a finite support and the true parameter is in the support, and thus their analysis requires some approximation scheme for dealing Bernoulli rewards.

5.4 Multiple-play Thompson Sampling Algorithm

Algorithm 8 is our MP-TS algorithm. While TS for single-play selects the top-1 arm based on a posterior sample $\widetilde{\mu}_i(t)$, MP-TS selects the top-L arms ranked by the posterior sample $\widetilde{\mu}_i(t)$. Like Kaufmann et al. [2012] and Agrawal and Goyal [2013a], we set the uniform prior on each arm.

In Section 5.3.3, we discussed that the necessary condition to achieve the optimal regret bound is to suppress the simultaneous draws of two or more suboptimal arms, which characterizes the difficulty of the MP bandit problem.

Note that it is easy to extend other asymptotically optimal SP bandit algorithms, such as KL-UCB, to the MP bandit problem. Nevertheless, we were not able to prove the optimality of these algorithms for the MP bandit problem though the achievability of the bound (5.4) on $N_i(T)$ is easily proved, and the simulation results in Section 5.7 also imply their achievability of the regret bound. This is because TS has quite a plausible property to suppress simultaneous draws as we discuss below.

Before the exact statement in the next section, we give an intuition for the natural extension of TS can have an asymptotically optimal regret bound in the MP bandit problem. Roughly speaking, a bandit algorithm with a logarithmic regret draws a suboptimal arm with probability $O(1/t)$ at the t-th round, which amounts to $O(\sum_{t=1}^{T} 1/t) = O(\log T)$ regret. Thus, two suboptimal arms are drawn at the same round with probability $O(1/t^2)$, which amounts to $O(\sum_{t=1}^{T} 1/t^2) = O(1)$ total simultaneous draws, provided that each suboptimal arm is selected independently.

In TS, the score $\widetilde{\mu}_i(t)$ for the choice of arms is generated randomly at each round from the posterior independently between each arm, which enables us to bound simultaneous draws as the above intuition. On the other hand, in KL-UCB (or in other index algorithms), the UCB score for the choice of arms is deterministic given the past results of rewards, which means that the scores of suboptimal arms may behave quite similarly in the worst case on the past rewards.

5.5 Asymptotically Optimal Regret Bound

In this section, we state the main theoretical result (Theorem 19). The analysis that leads to this theorem is discussed in Section 5.6.

Theorem 19. (Regret upper bound of MP-TS) *For any sufficiently small $\epsilon_1 > 0, \epsilon_2 > 0$, the regret of MP-TS is upper-bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \in [K] \setminus [L]} \left(\frac{(1+\epsilon_1)\Delta_{i,L} \log T}{d(\mu_i, \mu_L)} \right) + C_a(\epsilon_1, \mu_1, \mu_2, \ldots, \mu_K) + C_b(T, \epsilon_2, \mu_1, \mu_2, \ldots, \mu_K),$$

where, $C_a = C_a(\epsilon_1, \mu_1, \mu_2, \ldots, \mu_K)$ is a constant independent on T and is $O(\epsilon_1^{-2})$ when we regard $\{\mu_i\}_{i=1}^K$ as constants. The value $C_b = C_b(T, \epsilon_2, \mu_1, \mu_2, \ldots, \mu_K)$ is a function of T, which, by choosing proper ϵ_2, grows at a rate of $O(\log \log T)$.

By letting $\epsilon_1 = O((\log T)^{-1/3})$ we obtain

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \in [K] \setminus [L]} \frac{\Delta_{i,L} \log T}{d(\mu_i, \mu_L)} + O((\log T)^{2/3}) \tag{5.8}$$

and we see that MP-TS achieves the asymptotic bound in (5.7).

Expected regret and high-probability regret: Anantharam et al. [1987] originally derived a regret lower bound in a stronger form than (5.7) such that for any $\epsilon > 0$, the regret of a strongly consistent algorithm is lower-bounded as

$$\lim_{T \to \infty} \Pr \left[\frac{\mathrm{Reg}(T)}{\log T} \geq \sum_{i \in [K] \setminus [L]} \frac{(1-\epsilon)\Delta_{i,L}}{d(\mu_i, \mu_L)} \right] = 1.$$

Combining this with (5.8) we can easily see that MP-TS satisfies

$$\lim_{T \to \infty} \Pr \left[\frac{\mathrm{Reg}(T)}{\log T} \leq \sum_{i \in [K] \setminus [L]} \frac{(1+\epsilon)\Delta_{i,L}}{d(\mu_i, \mu_L)} \right] = 1, \tag{5.9}$$

that is, MP-TS is also asymptotically optimal in the sense of high probability. Since an algorithm satisfying (5.9) is not always asymptotically optimal in the sense of expectation, our result, an expected asymptotically optimal regret bound, is also stronger in this sense than the high-probability bound by Gopalan et al. [2014].

5.6 Regret Analysis

We first define some additional notation that are useful for our analysis in Section 5.6.1 then analyze the regret bound in Section 5.6.2. The proofs of all the lemmas, except for Lemma 20, are given in the Appendix.

5.6.1 Additional notation

Let $\mu_L^{(-)} = \mu_L - \delta$ and $\mu_i^{(+)} = \mu_i + \delta$ for $\delta > 0$ and $i \in [K] \setminus [L]$. We assume δ to be sufficiently small such that $\mu_L^{(-)} \in (\mu_{L+1}, \mu_L)$ and $\mu_i^{(+)} \in (\mu_i, \mu_L)$. We also define

$N_i^{\mathrm{suf}}(T) = \frac{\log T}{d(\mu_i^{(+)}, \mu_L^{(-)})}$. Intuitively, $N_i^{\mathrm{suf}}(T)$ is the sufficient number of explorations to make sure that arm i is not as good as arm L.

Events: let $\max_{i\in S}^{(m)} a_i$ denote the m-th largest element of $\{a_i\}_{i\in S} \in \mathbb{R}^{|S|}$, that is, $\max_{i\in S}{}^{(m)} a_i = \max_{S'\subset S:|S'|=m} \min_{i\in S'} a_i$. We define $\widetilde{\mu}^*(t) = \max_{i\in[K]}^{(L)} \widetilde{\mu}_i(t)$ as the L-th largest posterior sample at round t (i.e., the minimum posterior sample among the selected arms), and $\widetilde{\mu}^{**}_{\setminus i,j}(t) = \max_{k\in[K]\setminus\{i,j\}}^{(L-1)} \widetilde{\mu}_k(t)$ as the $(L-1)$-th largest posterior sample at round t except for arms i and j. Moreover, let $\nu = \frac{\mu_{L-1}+\mu_L}{2}$. Let us define the following events.

$$
\begin{aligned}
\mathcal{A}_i(t) &= \{i \in I(t)\},\\
\mathcal{B}(t) &= \{\widetilde{\mu}^*(t) \geq \mu_L^{(-)}\},\\
\mathcal{C}_i(t) &= \bigcap_{j\in[K]\setminus([L-1]\cup\{i\})} \{\widetilde{\mu}^{**}_{\setminus i,j}(t) \geq \nu\},\\
\mathcal{D}_i(t) &= \{N_i(t) < N_i^{\mathrm{suf}}(T)\}.
\end{aligned}
$$

Event $\mathcal{A}_i(t)$ states that arm i is sampled at round t, and $\mathcal{D}_i(t)$ states that arm i has not been sampled sufficiently yet. The complements of $\mathcal{B}(t)$ and $\mathcal{C}_i(t)$ are related to the underestimation of optimal arms. Since the optimal arms are sampled sufficiently, $\mathcal{B}^c(t)$ or $\mathcal{C}_i^c(t)$ should not occur very frequently.

5.6.2 Proof of Theorem 19

We first decompose the regret to the contribution of each arm. Recall that, the regret increase by drawing suboptimal arm i is determined by the optimal arm excluded in the selection set $I(t)$. Formally, for suboptimal arm i, let

$$
\Delta_i(t) = \begin{cases} (\max_{j\in[L]\setminus I(t)} \mu_j) - \mu_i & \text{if } I(t) \neq [L],\\ 0 & \text{otherwise,} \end{cases} \tag{5.10}
$$

and

$$
\mathrm{Reg}_i(T) = \sum_{t=1}^{T} \mathbf{1}\{i \in I(t)\}\Delta_i(t).
$$

From inequality (5.5) the following inequality is easily derived

$$
\mathrm{Reg}(T) \leq \sum_{i\in[K]\setminus[L]} \mathrm{Reg}_i(T).
$$

We next decompose $\mathrm{Reg}_i(T)$ into several terms by using events $\mathcal{A}$–$\mathcal{D}$. After giving bounds for these terms, we finally give the total regret bound, which proves Theorem 19. Note that, in bounding the deviation of Bernoulli means and Beta posteriors in the Appendix, our analysis borrowed some techniques developed in the context of the SP bandit problem, mostly from Agrawal and Goyal [2013a], and some from Honda and Takemura [2014].

Lemma 20. *The regret by drawing suboptimal arm $i > L$ is decomposed as*

$$\mathrm{Reg}_i(T) \le \underbrace{\sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}^c(t)\}}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i^c(t)\}}_{\text{(B)}} + \underbrace{\sum_{j \in [K]\setminus([L-1]\cup\{i\})} \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \mathcal{A}_j(t)\}}_{\text{(C)}} + \underbrace{\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{B}(t), \mathcal{D}_i^c(t)\}}_{\text{(D)}} + N_i^{\mathrm{suf}}(T)\Delta_{i,L},$$

where, for example, $\{\mathcal{A}, \mathcal{B}\}$ abbreviates $\{\mathcal{A} \cap \mathcal{B}\}$.

Roughly speaking,

- Term (A) corresponds to the case in which, some of the optimal arms are under-estimated.
- Term (B) corresponds to the case in which, arm i is selected and some of the arms in $[L-1]$ are under-estimated.
- Term (C) corresponds to the case in which, arm $i \in [K] \setminus [L]$ and $j \in [K] \setminus ([L-1] \cup \{i\})$ are simultaneously drawn. In particular, term (C) is unique in the MP bandit problem that causes additional regret increase, and in analyzing this term we fully use the fact that the samples of the posterior distributions on the arms are independent of each other.
- Term (D) corresponds to the case in which, arm i is selected after it is sufficiently explored.

Proof of Lemma 20. The contribution of suboptimal arm i to the regret is decomposed as follows. By using the fact $\Delta_i(t) \le 1$ and the following decomposition of an event

$$\begin{aligned}\mathcal{A}_i(t) &\subset \mathcal{B}^c(t) \cup \{\mathcal{A}_i(t), \mathcal{C}_i^c(t)\} \cup \{\mathcal{A}_i(t), \mathcal{B}(t), \mathcal{C}_i(t)\} \\ &\subset \mathcal{B}^c(t) \cup \{\mathcal{A}_i(t), \mathcal{C}_i^c(t)\} \cup \{\mathcal{A}_i(t), \mathcal{B}(t), \mathcal{D}_i^c(t)\} \cup \{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t)\},\end{aligned}$$

we have

$$\begin{aligned}\mathrm{Reg}_i(T) &= \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t)\}\Delta_i(t) \\ &\le \sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}^c(t)\} + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i^c(t)\} \\ &\quad + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{B}(t), \mathcal{D}_i^c(t)\} + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t)\}\Delta_i(t). \qquad (5.11)\end{aligned}$$

Recall that $\Delta_i(t)$ is defined as (5.10). At each round, when L and all suboptimal arms, except for i, are not selected, then $I(t) = \{1, 2, \ldots, L-1, i\}$; $\Delta_i(t) = \Delta_{i,L}$. Therefore,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t)\}\Delta_i(t) \\
&\leq \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t)\}\Delta_{i,L} + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \bigcup_{j \in [K]\setminus([L-1]\cup\{i\})} \mathcal{A}_j(t)\} \\
&\leq \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{D}_i(t)\}\Delta_{i,L} + \sum_{j \in [K]\setminus([L-1]\cup\{i\})} \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \mathcal{A}_j(t)\} \\
&\leq N_i^{\mathrm{suf}}(T)\Delta_{i,L} + \sum_{j \in [K]\setminus([L-1]\cup\{i\})} \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \mathcal{A}_j(t)\}. \qquad (5.12)
\end{aligned}$$

Summarizing (5.11) and (5.12) completes the proof. □

The following lemma bounds terms (A)–(D).

Lemma 21. (Bounds on individual terms) *Let $\epsilon_2 > 0$ be arbitrary. For sufficiently small δ and ϵ_2, the four terms are bounded in expectation as*

$$\mathbb{E}[(\mathrm{A})] = O\left(\frac{1}{(\mu_L - \mu_L^{(-)})^2}\right) = O\left(\frac{1}{\delta^2}\right), \qquad (5.13)$$

$$\mathbb{E}[(\mathrm{B})] = O(\log\log T), \qquad (5.14)$$

$$\mathbb{E}[(\mathrm{C})] \leq \sum_{j \in [K]\setminus([L-1]\cup\{i\})} \frac{\left(\epsilon_2 + 4T^{-\frac{\epsilon_2 \Delta_{L,L-1}^2}{8}}\right)\log T}{d(\mu_i, \mu_L)} + O(1),$$

and (5.15)

$$\mathbb{E}[(\mathrm{D})] \leq 2 + \frac{1}{d(\mu_i^{(+)}, \mu_i)} = O\left(\frac{1}{\delta^2}\right). \qquad (5.16)$$

The proof of Lemma 21 is in Section 5.11.2. Moreover, the following lemma bounds term (C) by choosing a proper value of ϵ_2.

Lemma 22. (Evaluation of ϵ_2-dependent factor) *By choosing an $O((\log\log T)/\log T)$ value of ϵ_2, we obtain $\mathbb{E}[(\mathrm{C})] = O(\log\log T)$.*

The proof of Lemma 22 is in Section 5.11.3. Now it suffices to evaluate $N_i^{\mathrm{suf}}(T) = \frac{\log T}{d(\mu_i^{(+)}, \mu_L^{(-)})}$ to complete the proof. From the convexity of KL divergence there exists a constant $c_i = c_i(\mu_i, \mu_L) > 0$ such that

$$d(\mu_i^{(+)}, \mu_L^{(-)}) = d(\mu_i + \delta, \mu_L - \delta) \geq (1 - c_i\delta)d(\mu_i, \mu_L)$$

and therefore

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}(T)] &\leq \sum_{i\in[K]\setminus[L]} \mathbb{E}[\mathrm{Reg}_i(T)] \leq \sum_{i\in[K]\setminus[L]} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t)\}\Delta_i(t)\right] \\ &\leq \sum_{i\in[K]\setminus[L]} \left\{\mathbb{E}\left[(\mathrm{A})+(\mathrm{B})+(\mathrm{C})+(\mathrm{D})\right] + N_i^{\mathrm{suf}}(T)\Delta_{i,L}\right\} \\ &\leq \underbrace{\sum_{i\in[K]\setminus[L]} \frac{\Delta_{i,L}\log T}{(1-c_i\delta)d(\mu_i,\mu_L)}}_{\text{main term}} + \underbrace{O\left(\frac{1}{\delta^2}\right)}_{C_a} + \underbrace{O(\log\log T)}_{C_b}.\end{aligned}$$

Since $(1-c_i\delta)^{-1} \leq 1+2c_i\delta$ for $c_i\delta \leq 1/2$, we complete the proof of Theorem 19 by letting $\epsilon_1 < 1/2$ and $\delta = \epsilon_1 / \max_{i\in[K]\setminus[L]} c_i = \Theta(\epsilon_1)$. □

5.7 Experiment

We ran a series of computer simulations[*2] to clarify the empirical properties MP-TS. The simulations involved the following three scenarios. In Scenarios 1 and 2, we used fixed arms similar to that of Garivier and Cappé [2011], and Scenario 3 is based on a click log dataset of advertisements on a commercial search engine.

Algorithms: the simulations involved MP-TS, Exp3.M [Uchiya et al., 2010], CUCB [Chen et al., 2013], and MP-KL-UCB. Exp3.M is a state-of-the-art adversarial bandit algorithm for the MP bandit problem[*3]. The learning rate γ of Exp3.M is set in accordance with Corollary 1 of Uchiya et al. [2010]. Note that the CUCB algorithm in the MP bandit problem at each round draws the top-L arms of the UCB indices $\widehat{\mu}_i + \sqrt{(3\log t)/(2N_i(t))}$. MP-KL-UCB is the algorithm that selects the top-L arms in accordance with the KL-UCB index $\sup_{q\in[\widehat{\mu}_i(t),1]} \{q | N_i(t) d(\widehat{\mu}_i(t), q) \leq \log t\}$.

Scenario 1 (5-armed bandits): the simulations include 5 Bernoulli arms with $\{\mu_1,\ldots,\mu_5\} = \{0.7, 0.6, 0.5, 0.4, 0.3\}$, and $L = 2$.

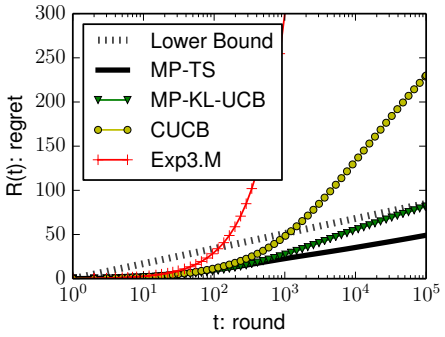
Scenario 2 (20-armed bandits): the simulations include 20 Bernoulli arms with $\mu_1 = 0.15$, $\mu_2 = 0.12$, $\mu_3 = 0.10$, $\mu_i = 0.05$ for $i \in \{4,5,\ldots,12\}$, $\mu_i = 0.03$ for $i \in \{13,14,\ldots,20\}$, and $L = 3$.

Scenario 3 (many-armed bandits, online advertisement based CTRs): we conducted another set of experiments with arms whose expectations were based on the dataset provided for KDD Cup[*4] 2012 track 2. The dataset involves a click log on soso.com (a large-scale search engine serviced by Tencent), which is composed of 149 million impressions (view of advertisements). We processed the data as follows. First, we excluded users of abnormally high click probability (i.e., users who had more than 1,000 impressions and

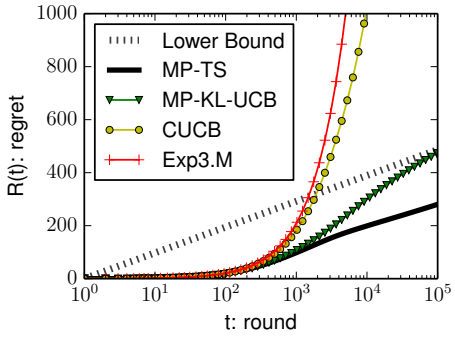
[*2] The source code of the simulations is available at https://github.com/jkomiyama/multiplaybanditlib.

[*3] Note that, Exp3.M is designed for the adversarial setting in which the rewards of arms are not necessarily i.i.d.

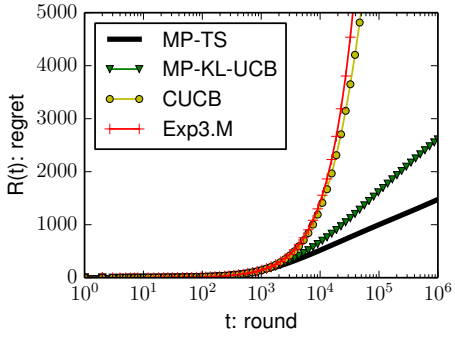
[*4] https://www.kddcup2012.org/



(a) Scenario 1



(b) Scenario 2



(c) Scenario 3

Fig. 5.2. Regret-round plots of algorithms. The regret in Scenarios 1 and 2 are averaged over 10,000 runs, and the regret in Scenario 3 is averaged over 1,000 runs. "Lower Bound" is the leading $\Omega(\log T)$ term of the RHS of inequality (5.7). We do not show Lower Bound in Scenario 3 because the coefficient of the bound can sometimes be quite large (i.e., in some runs, $1/d(\mu_{L+1}, \mu_L)$ is large).

more than 0.1 click probability) from the log. We also excluded minor advertisements (ads) that had less than 5,000 impressions. There are a wide variety of ads on a search engine (e.g., "rental cars", "music", etc.) and randomly picking ads from a search engine should yield a set of irrelevant ads. To address this issue, we selected popular queries that had more than 10^4 impressions and more than 50 ads that appeared on the query. As a result, 80 queries were obtained. The number of ads associated with each query ranged from 50 to 105, and the average click-through rate (CTR, the probability that the ad is clicked) of an ad on each query ranged from 1.15% to 6.86%. After that, each ad was converted into a Bernoulli arm with its expectations corresponding to the CTR of the ad. At the beginning of each run, one of the queries was randomly selected, and the bandit

simulation with the arms corresponding to the query and $L = 3$ is then conducted. This scenario was more difficult than the first two scenarios in the sense that 1) a larger number of arms were involved and 2) the reward gap among arms was very small.

The simulation results are shown in Figure 5.2. In all scenarios, the tendency is the same: our proposed MP-TS performs significantly better than the other algorithms. MP-KL-UCB is not as good as MP-TS, but clearly better than CUCB and Exp3.M. While it is unclear whether the slope of the regret of MP-KL-UCB converges to the asymptotic bound or not, the slope of the regret of TS quickly approaches the asymptotic lower bound.

5.7.1 Improvement of MP-TS based on the empirical means

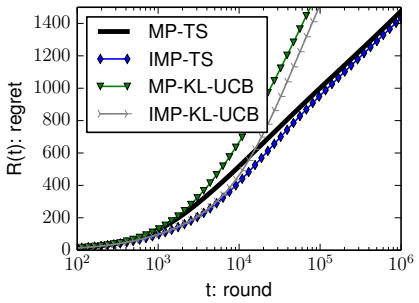


Fig. 5.3. Before/after comparison of MP-TS. All settings (except for algorithms) are the same as that of Scenario 3.

We now introduce an improved version of MP-TS (IMP-TS). In the theoretical analysis of the MP bandit problem, we observed that an extra loss arises when multiple suboptimal arms are drawn at the same round. Based on this observation, the new algorithm selects $L-1$ arms on the basis of empirical averages and selects the last arm on the basis of TS to avoid simultaneous draws of suboptimal arms. In other words, this algorithm is further aimed to minimize the regret by purely exploiting the knowledge in the top-$(L-1)$ arms; thus, limiting the exploration to only one arm. One might fear that this increase in exploitation could devastate the balance between exploration and exploitation. Although we provide no regret bound for the improved version of the algorithm, we expect that this algorithm will also achieve the asymptotic bound for the following reason. When we restrict the exploration to one arm, the number of opportunities for an arm to be explored may decrease, say, from T to T/L. Still, T/L opportunities are sufficient since $O(\log(T/L)) = O(\log T)$. In fact, the algorithm proposed by Anantharam et al. [1987] achieves the asymptotic bound even though $L-1$ arms are selected based on empirical means as in IMP-TS. Similarly, we define an improved version of MP-KL-UCB (IMP-KL-UCB) for selecting the first $L-1$ arms on the basis of empirical averages. The before/after analysis of this improvement is shown in Figure 5.3. One sees that, (i) MP-TS

still performs better than IMP-KL-UCB, and (ii) IMP-TS reduces the regret throughout the rounds. In particular, when the number of the rounds is small ($T \sim 10^3$–10^4), the advantage of IMP-TS is large.

5.8 Discussion

We extended TS to the multiple-play setting and proved its asymptotic optimality in terms of the regret. We considered the case in which the total reward is linear to the individual rewards of selected arms. The analysis here fully uses the independent property of posterior samples and paves the way to obtain a tight analysis on the multiple-play regret that depends on the combinatorial structure of arm selection. We now point out two promising directions for future work.

- **Position-dependent factors for online advertising:** it is well-known that the CTR of an ad is dependent on its position. Taking the position-dependent factor into consideration changes the MP bandit problem from the L-set selection problem to the L-sequence selection problem in which the position of L arms matters. For the starting point, we consider an extension of MP-TS for the cascade model Kempe and Mahdian [2008], Aggarwal et al. [2008] that corrects position-dependent bias in Section 5.10.
- **Non-Bernoulli distributions for general problems:** for the ease of argument, we exclusively consider the binary rewards. The analysis by Korda et al. [2013] is useful in extending our result to the case of the 1-d exponential families of rewards. Moreover, extending our result to multi-parameter reward distributions Burnetas and Katehakis [1996], Honda and Takemura [2014] is interesting.

5.9 Cases of Several Arms Having the Same Expectation

Up to now, we have assumed that all arms have distinct expectations. Here, we consider cases in which some arms have the same expectations. Without loss of generality, we assume $\mu_1 \geq \mu_2 \geq, \ldots, \geq \mu_K$. Let us call arms with a larger expectation than μ_L "strictly optimal" arms, arms with the same expectation as μ_L "marginal" arms, and arms with a smaller expectation than μ_L "strictly suboptimal" arms. Each arm is either strictly optimal, marginal, or strictly suboptimal.

Case 1: assume that all strictly optimal arms are distinct, that there is only one marginal arm, and that there are several strictly suboptimal arms with the same expectation. In this case, the regret bound of Theorem 19 holds because our analysis deals with each suboptimal arm separately.

Case 2: assume that there is only one marginal arm, that all strictly suboptimal arms are distinct, and that there are several strictly optimal arms with the same expectation. The regret bound also holds in this case since there is a gap between each strictly suboptimal arm and each strictly optimal arm.

Case 3: assume that all strictly optimal arms and strictly suboptimal arms are distinct and that there are several marginal arms with the same expectation. Unfortunately, we were unable to perform a meaningful analysis in this case. Intuitively, as stated by Agrawal and Goyal [2012] for the SP bandit, adding an additional marginal arm appears to require some extra exploration, which slightly increases the regret. However, the regret structure is more complex than the SP bandit because several marginal arms can be drawn simultaneously.

In summary, our Theorem 19 holds when the marginal arm is distinct. That is, $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{L-1} > \mu_L > \mu_{L+1} \geq \cdots \geq \mu_K$.

5.10 Cascade Model and Position-dependent MP Bandit Problem

Up to now, we assumed that the rewards of arms are independently and identically drawn from individual distributions. In this section, we relax this assumption and consider a wider class of the MP bandit problem. Remember that, one of our primary applications is multiple advertisement placement in the online advertising problem (c.f., Example 1). In this section, we interchangeably use the terms an advertisement (ad) and an arm. It is known that the CTR of an ad depends on the environment where the ad is placed, especially on the position of the ad. Among several models that explain this dependency on the position, the model that explains human behavior and agrees well with real data

Algorithm 9 Bias-Corrected Multiple-play Thompson sampling (BC-MP-TS) for binary rewards

Input: # of arms K, # of positions L, discount factors $\{\gamma_l(i)\}$
for $i = 1, 2, \ldots, K$ **do**
  $A_i, N_i \leftarrow 1, 2$
end for
$t \leftarrow 1$.
for $t = 1, 2 \ldots, T$ **do**
  for $i = 1, 2, \ldots, K$ **do**
    $B_i \leftarrow \max(N_i - A_i, 1)$
    $\widetilde{\mu}_i(t) \sim \mathrm{Beta}(A_i, B_i)$
  end for
  Select $I_l(t)$ $(l = 1, \ldots, L)$ in accordance with Section 5.10.2.
  for $l \in 1, 2, \ldots, L$ **do**
    if $\widehat{X}_i(t) = 1$ **then**
      $A_i \leftarrow A_i + 1$
    end if
    $N_i \leftarrow N_i + \prod_{l'=2}^{l} \gamma_{l'}(I_{l'-1}(t))$
  end for
end for

[Craswell et al., 2008] is the *cascade* model [Kempe and Mahdian, 2008, Aggarwal et al., 2008], with which it is assumed that the user scans the ads from top to bottom. Following Gatti et al. [2012], we define the discount factor $\gamma_l(i)$ for $l \geq 2$ as the probability that a user observing ad i in position $l-1$ will observe the ad in the next position. Namely, the MP bandit problem with a discount factor is defined as a MP bandit problem in which the arm at position l yields reward 1 with probability $\left(\prod_{l'=2}^{l} \gamma_{l'}(I_{l'-1}(t))\right) \mu_{I_l(t)}$, where $I_l(t)$ be the arm placed at the l-th position at round t. Note that, when we set $\gamma_l(i) = 1$ for any position $l \in [L]$ and ad i, this model is reduced to the model we have considered up to the previous section. In this case, the order of the L arms does not matter. Whereas, under a position-dependent discount factor smaller than 1, the order of L arms matters: the problem is not the selection of an L-set of arms, but an L-sequence of arms.

5.10.1 Thompson sampling for cascade model

In the cascade model, there is some probability that the arm at position $l > 1$ is not drawn. The probability that the arm at position l is drawn, $\prod_{l'=2}^{l} \gamma_{l'}(I_{l'-1}(t))$, can be considered as the *effective number of the draws* at position i. MP-TS (Algorithm 8) keeps A_i and B_i, which respectively correspond to the number of rewards 1 and 0. The number

of draws on the arm i is $N_i = A_i + B_i$. When we consider the cascade model, we need to take the effective number of draw into consideration. We introduce Bias-corrected MP-TS (BC-MP-TS, Algorithm 9). The crux of BC-MP-TS is that, for each arm that is selected, N_i should be increased not by 1, but by the effective number of draw for each position. Note that, when $\gamma_l(i) = 1$, BC-MP-TS is essentially the same as MP-TS.

5.10.2 Optimal arm selection and the regret

In general discount factor $\gamma_l(i)$, even if we have perfect information over the expectation of all arms $\{\mu_i\}_{i=1}^K$, the computation of the optimal sequence of L-arms at each round t (optimal arm selection) appears to be computationally intractable when K is large because we need to search all the possible allocation of K ads over L positions. Kempe and Mahdian [2008] proposed a polynomial-time approximation of the optimal arm selection. We can obtain the arm selection strategy for BC-MP-TS by using this approximation algorithm as an oracle and plugging $\{\widetilde{\mu}_i(t)\}_{i=1}^L$ as estimated expected rewards.

Ad-independent discount factor: when the discount factor is independent of the ad at that position (i.e., $\gamma_l(i) = \gamma_l$), the optimal arm selection is easy: just select μ_l (i.e., l-th optimal arm) on the l-th position. We define the arm selection strategy of BC-MP-TS as placing the arm of the l-th largest $\widetilde{\mu}_i$ (i.e., $I_l(t) = \max^{(l)}_{i\in[K]} \widetilde{\mu}_i$) on the l-th position.

Regret: naturally, the regret per round is defined as the difference between the expected reward of the optimal arm selection and that of an algorithm. Namely,

$$\mathrm{Reg}(T) = \sum_{t=1}^{T}\sum_{l=1}^{L}\left(\prod_{l'=2}^{l}\gamma_{l'}(I_{\mathrm{opt}}(l'-1))\mu_{I_{\mathrm{opt}}(l)} - \underbrace{\prod_{l'=2}^{l}\gamma_{l'}(I_{l'-1}(t))}_{\text{effective number of draw at position } l}\times\mu_{I_l(t)}\right),$$

where $(I_{\mathrm{opt}}(1), \ldots, I_{\mathrm{opt}}(L))$ is the optimal arm selection. In the case of the ad-independent discount factor, we conjecture that the asymptotic regret lower bound should be identical to the case of no-discount factor that we have analyzed (i.e., inequality (5.7)). Although we do not prove any regret bound for this cascade model, the conjecture is supported by the fact that (i) by identifying the top-L arm we immediately obtain the optimal arm selection, (ii) algorithms should require $\log T/d(\mu_i, \mu_L)$ number of effective draws to convince that suboptimal arm $i > L$ is not as good as arm L, and (iii) the best situation is that the simultaneous draw of several optimal arms rarely occurs: arm L is pushed out instead of arm i, and the regret increase per an effective draw is $\mu_L - \mu_i$. In the case of the general discount factor, the problem is subtler because a slight difference in $\{\mu_i\}$ can change the optimal arm selection.

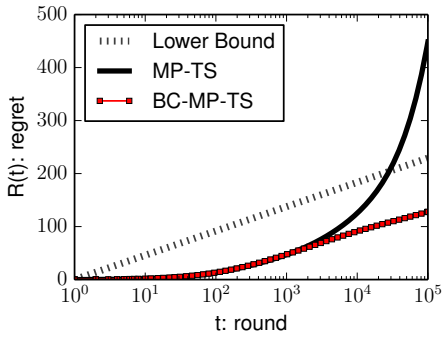


Fig. 5.4. Simulation with a discount factor. Lower Bound is the leading $\Omega(\log T)$ term of the RHS of inequality (5.7), which we have conjectured to be the lower bound for the cascade model with the ad-independent discount factor in Section 5.10.2. The regret is averaged over $10,000$ runs.

5.10.3 Experiment of cascade model

This simulation adapts the cascade model and involves a constant discount factor $\gamma_l(i) = 0.7$ for any position and arm. There are 9 Bernoulli arms with $\mu_1 = 0.24, \mu_2 = 0.21, \ldots, \mu_9 = 0.00$ and $L = 3$. In this case the optimal arm selection strategy is to choose $\{I_1(t), I_2(t), I_3(t)\} = \{\mu_1, \mu_2, \mu_3\}$ (c.f., Section 5.10.2). The regret of the algorithms is shown in 5.4. On one hand, MP-TS failed to have a small regret due to its ignorance to the discount factors. On the other hand, the slope of BC-MP-TS quickly approaches the conjectured Lower Bound, which is empirical evidence of the ability of BC-MP-TS to correct the position-dependent bias.

5.11 Proofs

5.11.1 Lemmas

Lemma 23. (Lemma 2 in Agrawal and Goyal [2013a]) *Let* $k \in [K]$, $n \geq 0$ *and* $x < \mu_k$. *Let* $\widehat{\mu}_{k,n}$ *be the empirical average of* n *samples from* Bernoulli(μ_k). *Let* $p_{k,n}(x) = 1 - F^{\mathrm{beta}}_{\widehat{\mu}_{k,n}n+1,(1-\widehat{\mu}_{k,n})n+1}(y)$ *be the probability that the posterior sample from the Beta distribution with its parameter* $\widehat{\mu}_{k,n}n+1, (1-\widehat{\mu}_{k,n})n+1$ *exceeds* x. *Then, its average over*

runs is bounded as

$$\mathbb{E}\left[\frac{1}{p_{k,n}(x)}\right] \leq \begin{cases} 1 + \frac{3}{\Delta_k(x)} & (n < 8/\Delta_k(x)) \\ 1 + \Theta\Bigg(e^{-\Delta_k(x)^2 n/2} + \frac{1}{(n+1)\Delta_k(x)^2} e^{-D_k(x)n} \\ \qquad + \frac{1}{e^{\Delta_k(x)^2 n/4} - 1}\Bigg) & (n \geq 8/\Delta_k(x)), \end{cases}$$

where $\Delta_k(x) = \mu_k - x, D_k(x) = d(x, \mu_k)$.

In the proof of Lemma 21 we use the following Lemmas 24, 25, and 26 several times. Lemma 24 is essentially the combination of the existing techniques of Agrawal and Goyal [2013a] and Honda and Takemura [2014]. Lemmas 25 and 26 are also existing techniques that appear in several previous analyses in Bayesian bandits with Bernoulli arms.

Lemma 24. *Let* $k \in [K]$, $z < \mu_k$ *be arbitrary,* $\mathcal{S}(t)$, $\mathcal{T}(t)$, *and* $\mathcal{U}(t)$ *be events such that*

(i) if $\{\widetilde{\mu}_k(t) \geq z\}$, $\mathcal{S}(t)$, *and* $\mathcal{T}(t)$ *occurred then the arm* k *is drawn at round* t,
(ii) $\widetilde{\mu}_k(t)$, $\mathcal{S}(t)$ *and* $\mathcal{T}(t)$ *are mutually independent given* $\{\widehat{\mu}_i(t)\}_{i=1}^K$ *and* $\{N_i(t)\}_{i=1}^K$.
(iii) The event $\mathcal{U}(t)$ *is deterministic given* $\{\widehat{\mu}_i(t)\}_{i=1}^K$ *and* $\{N_i(t)\}_{i=1}^K$.
(iv) Given $\{\widehat{\mu}_i(t)\}_{i=1}^K$ *and* $\{N_i(t)\}_{i=1}^K$ *such that* $\mathcal{U}(t)$ *holds,* $\mathcal{T}(t)$ *occurs with probability at least* $q > 0$.

Then

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t), \mathcal{U}(t), N_k(t) < N_c\}\right] = O\left(\frac{1}{q(\mu_k - z)^2}\right) + N_c \frac{1-q}{q}.$$

In particular, by setting $\mathcal{T}(t)$ *and* $\mathcal{U}(t)$ *the trivial events that always hold* ($q = 1$)*, we obtain the following inequality:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t)\}\right] = O\left(\frac{1}{(\mu_k - z)^2}\right). \tag{5.17}$$

Proof. First we have

$$\begin{aligned} &\sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t), \mathcal{U}(t), N_k(t) < N_c\} \\ &\leq \sum_{n=0}^{N_c} \sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t), \mathcal{U}(t), N_k(t) = n\} \\ &\leq \sum_{n=0}^{N_c} \sum_{m=1}^{T} \mathbf{1}\left[m \leq \sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t), \mathcal{U}(t), N_k(t) = n\}\right]. \end{aligned} \tag{5.18}$$

Here note that the event

$$m \leq \sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t), \mathcal{U}(t), N_k(t) = n\}$$

implies that the event

$$\{\mathcal{S}(t), \mathcal{U}(t), N_k(t) = n\} \tag{5.19}$$

occurred for at least m rounds and $\{\widetilde{\mu}_k(t) < z\}$ or $\mathcal{T}^c(t)$ occurred for the first m rounds such that (5.19) occurred. Thus, by using the mutual independence of $\{\widetilde{\mu}_k(t) < z\}$, $\mathcal{S}(t)$, and $\mathcal{T}(t)$, we have

$$\Pr\left[m \leq \sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t), \mathcal{U}(t), N_k(t) = n\} \middle| \widehat{\mu}_{k,n}\right] \leq (1 - p_{k,n}(z)q)^m \tag{5.20}$$

and therefore

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t), \mathcal{U}(t), N_k(t) < N_c\} \middle| \widehat{\mu}_{k,n}\right] &\leq \sum_{n=0}^{N_c} \sum_{m=1}^{T} (1 - p_{k,n}(z)q)^m \quad \text{(by (5.18) and (5.20))} \\
&\leq \sum_{n=0}^{N_c} \frac{1 - p_{k,n}(z)q}{p_{k,n}(z)q} = \frac{1}{q} \sum_{n=0}^{T-1} \left(\frac{1}{p_{k,n}(z)} - 1\right) + N_c \frac{1-q}{q}.
\end{aligned}$$

By using Lemma 23, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n=0}^{T-1} \left(\frac{1}{p_{k,n}(z)} - 1\right)\right] \\
&\leq \frac{24}{\Delta_k(z)^2} + \sum_{n=\lceil 8/\Delta_k(z) \rceil}^{T-1} O\left(e^{-\Delta_k(z)^2 n/2} + \frac{e^{-D_k(z)n}}{(n+1)\Delta_k(z)^2} + \frac{1}{e^{\Delta_k(z)^2 n/4} - 1}\right).
\end{aligned} \tag{5.21}$$

By using the fact that $D_k(z) = d(z, \mu_k) = \Omega(1/(\mu_k - z)^2)$ (from the Pinsker's inequality), it is easy to verify that the RHS of (5.21) is $O(1/(\mu_k - z)^2)$. By using these facts, we finally obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}_k(t) < z, \mathcal{S}(t), \mathcal{U}(t), N_k(t) < N_c\}\right] &\leq \frac{1}{q}\mathbb{E}\left[\sum_{n=0}^{T-1} \left(\frac{1}{p_{k,n}(z)} - 1\right)\right] + N_c \frac{1-q}{q} \\
&= O\left(\frac{1}{q(\mu_k - z)^2}\right) + N_c \frac{1-q}{q},
\end{aligned}$$

which concludes the proof of the lemma. □

Lemma 25. (Deviation of empirical averages, Agrawal and Goyal [2013a, Appendix B.1]) *Let $k \in [K]$ and $z > \mu_k$ be arbitrary. Then,*

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \mathbf{1}\{\mathcal{A}_k(t), \widehat{\mu}_k(t) > z\}\right] < 1 + \frac{1}{d(z, \mu_k)}.$$

Lemma 26. (Deviation of Beta posteriors) *Let $k \in [K]$, $x_1, x_2 \in [0, 1]$ be arbitrary values such that $x_1 > x_2$, and $n \geq 1$. Then,*

$$\mathbb{P}(\widetilde{\mu}_k(t) \geq x_1 | \widehat{\mu}_k(t) \leq x_2, N_k(t) = n) \leq \exp\left(-d(x_2, x_1)n\right).$$

Proof. Note that, this lemma is essentially the same as the first display in Agrawal and Goyal [2013a, Appendix B.2]. While Agrawal and Goyal [2013a] provide a bound for $N_k(t) > n$, the bound in our lemma is for $N_k(t) = n$. For the sake of rigor, we write the proof here.

$$
\begin{aligned}
&\mathbb{P}(\widetilde{\mu}_j(t) \geq x_1 | \widehat{\mu}_j(t) \leq x_2, N_j(t) = n) \\
&= \mathbb{P}\left(\widetilde{\mu} \sim \mathrm{Beta}(\widehat{\mu}_j(t)n + 1, (1 - \widehat{\mu}_j(t))n + 1), \widetilde{\mu} \geq x_1 \middle| \widehat{\mu}_j(t) \leq x_2 \right) \\
&= 1 - F^{\mathrm{beta}}_{x_2 n+1, (1-x_2)n+1}(x_1) \\
&= F^{\mathrm{B}}_{n+1, x_1}(x_2 n) \\
&\qquad\qquad \text{(by the Beta-Binomial equality)} \\
&\leq F^{\mathrm{B}}_{n, x_1}(x_2 n) \leq \exp\left(-d(x_2, x_1)n\right) \\
&\qquad\qquad \text{(by the Chernoff bound).}
\end{aligned}
$$

□

5.11.2 Proof of Lemma 21

Evaluation of term (A):

Proof. Here, we prove inequality (5.13). Recall that

$$
(\mathrm{A}) = \sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}^c(t)\} = \sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}^*(t) < \mu_L^{(-)}\}.
$$

Since $\widetilde{\mu}^*(t)$ is the L-th largest posterior sample among arms at round t, $\widetilde{\mu}^*(t) < \mu_L^{(-)}$ implies that, there exists at least one arm in $[L]$ with its posterior sample smaller than $\mu_L^{(-)}$. Namely,

$$
\{\widetilde{\mu}^*(t) < \mu_L^{(-)}\} \subset \bigcup_{k \in [L]} \{\widetilde{\mu}_k(t) < \mu_L^{(-)}\},
$$

and therefore

$$
\begin{aligned}
&\{\widetilde{\mu}^*(t) < \mu_L^{(-)}\} \\
&= \bigcup_{k \in [L]} \{\widetilde{\mu}_k(t) < \mu_L^{(-)}, \widetilde{\mu}^*(t) < \mu_L^{(-)}\} \\
&= \bigcup_{k \in [L]} \{\widetilde{\mu}_k(t) < \mu_L^{(-)}, \max_{j \in [L]}{}^{(L)} \widetilde{\mu}_j(t) < \mu_L^{(-)}\} \\
&\subset \bigcup_{k \in [L]} \{\widetilde{\mu}_k(t) < \mu_L^{(-)}, \max_{j \in [L] \setminus \{k\}}{}^{(L)} \widetilde{\mu}_j(t) < \mu_L^{(-)}\}.
\end{aligned}
$$

By using the union bound, we obtain

$$\mathbf{1}\{\widetilde{\mu}^*(t) < \mu_L^{(-)}\} \leq \sum_{k\in[L]} \mathbf{1}\{\widetilde{\mu}_k(t) < \mu_L^{(-)}, \max_{j\in[L]\setminus\{k\}}^{(L)} \widetilde{\mu}_j(t) < \mu_L^{(-)}\}.$$

Note that the event $\max_{j\in[L]\setminus\{k\}}^{(L)} \widetilde{\mu}_j(t) < \mu_L^{(-)}$ satisfies the condition for the event $\mathcal{S}(t)$ in (5.17) in Lemma 24 with $z := \mu_L^{(-)}$. Therefore we obtain from Lemma 24 that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\widetilde{\mu}^*(t) < \mu_L^{(-)}\}\right] = O\left(\frac{1}{(\mu_k - \mu_L^{(-)})^2}\right) = O\left(\frac{1}{(\mu_L - \mu_L^{(-)})^2}\right),$$

which concludes the proof of inequality (5.13). □

Evaluation of term (B):

Proof. Here, we prove inequality (5.14). We have,

$$\begin{aligned}
(\mathrm{B}) &= \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i^c(t)\} \\
&= \sum_{t=1}^{T} \mathbf{1}\left\{\bigcup_{j\in[K]\setminus([L-1]\cup\{i\})} \{\mathcal{A}_i(t), \widetilde{\mu}_{\setminus i,j}^{**}(t) < \nu\}\right\} \\
&= \sum_{t=1}^{T} \sum_{j\in[K]\setminus([L-1]\cup\{i\})} \mathbf{1}\left\{\mathcal{A}_i(t), \widetilde{\mu}_{\setminus i,j}^{**}(t) < \nu\right\} \\
&= \sum_{t=1}^{T} \sum_{j\in[K]\setminus([L-1]\cup\{i\})} \left\{\mathbf{1}\left\{\mathcal{A}_i(t), \widehat{\mu}_i(t) > \mu_L\right\} + \mathbf{1}\left\{\mathcal{A}_i(t), \widehat{\mu}_i(t) \leq \mu_L, \widetilde{\mu}_{\setminus i,j}^{**}(t) < \nu\right\}\right\}.
\end{aligned} \tag{5.22}$$

In the following, we bound the first and the second terms in the inner sum of the last line of (5.22). From Lemma 25, the first term of (5.22) is bounded as

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left\{\mathcal{A}_i(t), \widehat{\mu}_i(t) > \mu_L\right\}\right] \leq 1 + \frac{1}{d(\mu_L, \mu_i)} = O(1).$$

On the other hand, the second term of (5.22) is transformed as

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbf{1}\left\{\mathcal{A}_i(t), \widehat{\mu}_i(t) \leq \mu_L, \widetilde{\mu}_{\setminus i,j}^{**}(t) < \nu\right\} \\
&\leq \frac{\log\log T}{d(\mu_L, \nu)} + \sum_{t=1}^{T} \mathbf{1}\left\{\mathcal{A}_i(t), N_i(t) > \frac{\log\log T}{d(\mu_L, \nu)}, \widehat{\mu}_i(t) \leq \mu_L, \widetilde{\mu}_{\setminus i,j}^{**}(t) < \nu\right\} \\
&\leq \frac{\log\log T}{d(\mu_L, \nu)} + \sum_{t=1}^{T} \mathbf{1}\left\{N_i(t) > \frac{\log\log T}{d(\mu_L, \nu)}, \widehat{\mu}_i(t) \leq \mu_L, \widetilde{\mu}_{\setminus i,j}^{**}(t) < \nu\right\}.
\end{aligned}$$

Since $\widetilde{\mu}_{\setminus i,j}^{**}(t)$ is the $(L-1)$-th largest posterior sample among arms except for arms i and j, $\widetilde{\mu}_{\setminus i,j}^{**}(t) < \nu$ indicates that, the number of arms excluding i and j with posterior

samples larger than or equal to ν is at most $L-2$, and thus at least one arm among $[L-1]$ has its posterior smaller than ν. Namely,

$$\begin{aligned}\{\widetilde{\mu}^{**}_{\setminus i,j}(t) < \nu\} &= \{\max_{l\in[K]\setminus\{i,j\}}{}^{(L-1)}\widetilde{\mu}_l(t) < \nu\} \\ &= \bigcup_{k\in[L-1]} \{\widetilde{\mu}_k(t) < \nu, \max_{l\in[K]\setminus\{i,j\}}{}^{(L-1)}\widetilde{\mu}_l(t) < \nu\} \\ &\subset \bigcup_{k\in[L-1]} \{\widetilde{\mu}_k(t) < \nu, \max_{l\in[K]\setminus\{i,j,k\}}{}^{(L-1)}\widetilde{\mu}_l(t) < \nu\}.\end{aligned}$$

By using this, we have

$$\begin{aligned}&\sum_{t=1}^{T} \mathbf{1}\left\{N_i(t) > \frac{\log\log T}{d(\mu_L,\nu)}, \widehat{\mu}_i(t) \le \mu_L, \widetilde{\mu}^{**}_{\setminus i,j}(t) < \nu\right\} \\ &\le \sum_{t=1}^{T}\sum_{k\in[L-1]} \mathbf{1}\Big\{N_i(t) > \frac{\log\log T}{d(\mu_L,\nu)}, \widehat{\mu}_i(t) \le \mu_L, \widetilde{\mu}_k(t) < \nu, \max_{l\in[K]\setminus\{i,j,k\}}{}^{(L-1)}\widetilde{\mu}_l(t) < \nu\Big\}.\end{aligned}$$

Moreover, let $\nu_2 = (\nu+\mu_L)/2 = (\mu_{L-1}+3\mu_L)/4$. For $k \in [L-1]$, $\mu_k > \nu > \nu_2 > \mu_L$ and

$$\begin{aligned}&\mathbb{P}\{\widetilde{\mu}_k(t) < \nu, N_k(t) \ge \frac{\log T}{2(\nu-\nu_2)^2}\} \\ &\le \sum_{n=\frac{\log T}{2(\nu-\nu_2)^2}}^{T} \mathbb{P}\{\widetilde{\mu}_k(t) < \nu, N_k(t) = n\} \\ &\le \sum_{n=\frac{\log T}{2(\nu-\nu_2)^2}}^{T} \mathbb{P}\{\widetilde{\mu}_k(t) < \nu, \widehat{\mu}_k(t) > \nu_2, N_k(t) = n\} + \sum_{n=\frac{\log T}{2(\nu-\nu_2)^2}}^{T} \mathbb{P}\{\widehat{\mu}_k(t) \le \nu_2, N_k(t) = n\} \\ &\le \sum_{n=\frac{\log T}{2(\nu-\nu_2)^2}}^{T} \mathrm{e}^{-d(\nu_2,\nu)n} + \sum_{n=\frac{\log T}{2(\nu-\nu_2)^2}}^{T} \mathbb{P}\{\widehat{\mu}_k(t) \le \nu_2, N_k(t) = n\} \text{ (by Lemma 26)} \\ &\le \sum_{n=\frac{\log T}{2(\nu-\nu_2)^2}}^{T} \mathrm{e}^{-d(\nu_2,\nu)n} + \sum_{n=\frac{\log T}{2(\nu-\nu_2)^2}}^{T} \mathrm{e}^{-d(\nu_2,\mu_k)n} \text{ (by Chernoff bound)} \\ &= O(1/T) \text{ (by } (\mu_k-\nu_2) > (\nu-\nu_2) \text{ and Pinsker's inequality)}\end{aligned}$$

and thus

$$\sum_{t=1}^{T}\sum_{k\in[L-1]}\mathbf{1}\Big\{N_i(t)>\frac{(\log T)^{2/3}}{d(\mu_L,\nu)},\widehat{\mu}_i(t)\le\mu_L,\widetilde{\mu}_k(t)<\nu,\max_{l\in[K]\setminus\{i,j,k\}}{}^{(L-1)}\widetilde{\mu}_l(t)<\nu\Big\}$$
$$\le\sum_{t=1}^{T}\sum_{k\in[L-1]}\mathbf{1}\Big\{N_i(t)>\frac{(\log T)^{2/3}}{d(\mu_L,\nu)},N_k(t)<\frac{\log T}{2(\nu-\nu_2)^2},\widehat{\mu}_i(t)\le\mu_L,\widetilde{\mu}_k(t)<\nu,\max_{l\in[K]\setminus\{i,j,k\}}{}^{(L-1)}\widetilde{\mu}_l(t)<\nu\Big\}$$
$$+\sum_{t=1}^{T}\sum_{k\in[L-1]}\mathbf{1}\Big\{\widetilde{\mu}_k(t)<\nu,N_k(t)\ge\frac{\log T}{2(\nu-\nu_2)^2}\Big\}$$
$$\le\sum_{t=1}^{T}\sum_{k\in[L-1]}\mathbf{1}\Big\{N_i(t)>\frac{(\log T)^{2/3}}{d(\mu_L,\nu)},N_k(t)<\frac{\log T}{2(\nu-\nu_2)^2},\widehat{\mu}_i(t)\le\mu_L,\widetilde{\mu}_k(t)<\nu,\max_{l\in[K]\setminus\{i,j,k\}}{}^{(L-1)}\widetilde{\mu}_l(t)<\nu\Big\}$$
$$+O(1).$$

Here, $z:=\nu$, $\mathcal{S}(t):=\{\max^{(L-1)}_{l\in[K]\setminus\{i,j,k\}}\widetilde{\mu}_l(t)<\nu\}$, $\mathcal{T}(t):=\{\widetilde{\mu}_i(t)\le\nu\}$, and $\mathcal{U}(t):=\{\widehat{\mu}_i(t)\le\mu_L\}$ satisfy the conditions in Lemma 24. Under $\mathcal{U}(t)$, $\mathcal{T}(t)$ holds with probability at least

$$1-\exp\left(-d(\mu_L,\nu)\left(\frac{\log\log T}{d(\mu_L,\nu)}\right)\right)=1-(\log T)^{-1}$$

by Lemma 26. Therefore, by using Lemma 24 with $N_c=\log T/(2(\nu-\nu_2)^2)$, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\Big\{N_i(t)>\frac{\log\log T}{d(\mu_L,\nu)},N_k(t)<\frac{\log T}{2(\nu-\nu_2)^2},\widehat{\mu}_i(t)\le\mu_L,\widetilde{\mu}_k(t)<\nu,\max_{l\in[K]\setminus\{i,j,k\}}{}^{(L-1)}\widetilde{\mu}_l(t)<\nu\Big\}\right]$$
$$\le O\left(\frac{1}{(1-(\log T)^{-1})(\mu_k-\nu)^2}\right)+O\left(\frac{(\log T)^{-1}}{1-(\log T)^{-1}}\frac{\log T}{2(\nu-\nu_2)^2}\right)=O(1).\quad(5.23)$$

From (5.23) and the union bound over $k\in[L-1]$, the second term of (5.22) is $O(1)$. In summary, term (B) is $O(\log\log T)$ in expectation. □

Evaluation of term (C):

Proof. Here, we prove inequality (5.15). Recall that,

$$(\mathrm{C})=\sum_{j\in[K]\setminus([L-1]\cup\{i\})}\sum_{t=1}^{T}\mathbf{1}\{\mathcal{A}_i(t),\mathcal{A}_j(t),\mathcal{C}_i(t),\mathcal{D}_i(t)\}.$$

Remember that $\nu_2=(\nu+\mu_L)/2=(\mu_{L-1}+3\mu_L)/4$. Note that, we defined ν and ν_2 such that $\mu_{L-1}>\nu>\nu_2>\mu_L$, $O(\mu_{L-1}-\nu)=O(\nu-\nu_2)=O(\nu_2-\mu_L)=O(\mu_{L-1}-\mu_L)=O(1)$

as a function of T. Then,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}_i(t), \mathcal{D}_i(t)\} \\
&= \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \widehat{\mu}_j(t) > \nu_2\} + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \widehat{\mu}_j(t) \leq \nu_2\} \\
&\leq \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_j(t), \widehat{\mu}_j(t) > \nu_2\} + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \widehat{\mu}_j(t) \leq \nu_2\}.
\end{aligned} \tag{5.24}$$

By using Lemma 25 with $z := \nu_2$, the first term in (5.24) is bounded as

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_j(t), \widehat{\mu}_j(t) > \nu_2\}\right] &\leq 1 + \frac{1}{d(\nu_2, \mu_j)} \\
&= O\left(\frac{1}{(\nu_2 - \mu_j)^2}\right) = O\left(\frac{1}{(\mu_{L-1} - \mu_L)^2}\right) = O(1).
\end{aligned}$$

We now bound the second term in (5.24). Let $\mathcal{C}'_{i,j}(t) = \{\widetilde{\mu}^{**}_{\setminus i,j}(t) \geq \nu\} \supset \mathcal{C}_i(t)$. Let $\mathcal{E}_j(t) = \{N_j(t) \geq \epsilon_2 \log T\}$. We have,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \widehat{\mu}_j(t) \leq \nu_2\} \\
&\leq \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}'_{i,j}(t), \mathcal{D}_i(t), \widehat{\mu}_j(t) \leq \nu_2\} \\
&\leq \epsilon_2 \log T + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}'_{i,j}(t), \mathcal{D}_i(t), \widehat{\mu}_j(t) \leq \nu_2, \mathcal{E}_j(t)\}. \\
&\leq \epsilon_2 \log T + \sum_{n=0}^{N_i^{\mathrm{suf}}(T)-1} \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}'_{i,j}(t), N_i(t) = n, \widehat{\mu}_j(t) \leq \nu_2, \mathcal{E}_j(t)\}.
\end{aligned}$$

In the following, we bound

$$\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{A}_j(t), \mathcal{C}'_{i,j}(t), N_i(t) = n, \widehat{\mu}_j(t) \leq \nu_2, \mathcal{E}_j(t)\}. \tag{5.25}$$

Note that, (5.25) is at most 1 since $\{\mathcal{A}_i(t), N_i(t) = n\}$ occurs at most once. Let τ be the first round (if exists) at which $\{\mathcal{C}'_{i,j}(t), \widetilde{\mu}^{**}_{\setminus i,j}(t) \leq \widetilde{\mu}_i(t), \mathcal{A}_i(t), N_i(t) = n\}$ is satisfied. It is necessary that $\{\widetilde{\mu}_j(\tau) \geq \widetilde{\mu}^{**}_{\setminus i,j}(\tau)\}$ for (5.25) to be 1: this is because, (i) both $\widetilde{\mu}_i(\tau)$ and $\widetilde{\mu}_j(\tau)$ need to be larger than $\widetilde{\mu}^{**}_{\setminus i,j}(\tau)$ for the simultaneous draw of arms i and j, (ii) and if $\widetilde{\mu}_j(\tau) < \widetilde{\mu}^{**}_{\setminus i,j}(\tau)$ then arm i is drawn and thus $\{N_i(t) = n\}$ is never satisfied after $t > \tau$. Here,

$$\mathbb{P}\{\widetilde{\mu}_j(\tau) \geq \widetilde{\mu}^{**}_{\setminus i,j}(\tau), \widetilde{\mu}^{**}_{\setminus i,j}(\tau) \geq \nu, \widehat{\mu}_j(\tau) \leq \nu_2\} \leq \exp\left(-d(\nu_2, \nu) N_j(\tau)\right),$$

by Lemma 26. Therefore, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{A}_i(t),\mathcal{A}_j(t),\mathcal{C}_i(t),N_i(t)=n,\widehat{\mu}_j(t)\le\nu_2\}\right] \le \exp\left(-d(\nu_2,\nu)\epsilon_2\log T\right)=T^{-\epsilon_2 d(\nu_2,\nu)}.$$

In summary, the second term in (5.24) is bounded as

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{A}_i(t),\mathcal{A}_j(t),\mathcal{C}_i(t),\mathcal{D}_i(t),\widehat{\mu}_j(t)\le\nu_2\}\right]\\ &\le \epsilon_2\log T+N_i^{\mathrm{suf}}(T)T^{-\epsilon_2 d(\nu_2,\nu)}\\ &\le\left(\epsilon_2+\frac{4T^{-\epsilon_2 d(\nu_2,\nu)}}{d(\mu_i,\mu_L)}\right)\log T \qquad (\text{by } (1+\delta)^2<4),\end{aligned}$$

and thus,

$$\begin{aligned}&\mathbb{E}[(\mathrm{C})]\\ &\le\sum_{j\in[K]\setminus([L-1]\cup\{i\})}\left(\frac{\left(\epsilon_2+4T^{-\epsilon_2 d(\nu_2,\nu)}\right)\log T}{d(\mu_i,\mu_L)}\right)+O(1)\\ &\le\sum_{j\in[K]\setminus([L-1]\cup\{i\})}\left(\frac{\left(\epsilon_2+4T^{-\epsilon_2\Delta_{L,L-1}^2/8}\right)\log T}{d(\mu_i,\mu_L)}\right)+O(1),\end{aligned}$$

where we used the fact that $d(\nu_2,\nu)\ge 2(\nu-\nu_2)^2=2\times((\mu_{L-1}-\mu_L)/4)^2$ in the last transformation. □

Evaluation of term (D):

Proof. Here, we prove inequality (5.16). We first divide term (D) into two subterms as

$$\begin{aligned}\mathbb{E}[(\mathrm{D})]&=\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{A}_i(t),\mathcal{B}(t),N_i(t)\ge N_i^{\mathrm{suf}}(T)\}\right]\\ &\le\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{A}_i(t),\mathcal{B}(t),\widehat{\mu}_i(t)>\mu_i^{(+)},N_i(t)\ge N_i^{\mathrm{suf}}(T)\}\right]\\ &+\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{A}_i(t),\mathcal{B}(t),\widehat{\mu}_i(t)\le\mu_i^{(+)},N_i(t)\ge N_i^{\mathrm{suf}}(T)\}\right].\end{aligned}\tag{5.26}$$

On one hand, the first term in (5.26) is bounded as

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{A}_i(t),\mathcal{B}(t),\widehat{\mu}_i(t)>\mu_i^{(+)},N_i(t)\ge N_i^{\mathrm{suf}}(T)\}\right]\\ &\le\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{A}_i(t),\widehat{\mu}_i(t)>\mu_i^{(+)}\}\right]\\ &\le 1+\frac{1}{d(\mu_i^{(+)},\mu_i)} \qquad (\text{by Lemma 25}).\end{aligned}$$

On the other hand, each component of the second term of (5.26) is bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{1}[\mathcal{A}_i(t), \mathcal{B}(t), \widehat{\mu}_i(t) \leq \mu_i^{(+)}, N_i(t) \geq N_i^{\mathrm{suf}}(T)]\right] \\
&\leq \mathbb{E}\left[\mathbf{1}[\widetilde{\mu}_i(t) \geq \mu_L^{(-)}, \widehat{\mu}_i(t) \leq \mu_i^{(+)}, N_i(t) \geq N_i^{\mathrm{suf}}(T)]\right] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\mathbf{1}[\widetilde{\mu}_i(t) \geq \mu_L^{(-)}, \widehat{\mu}_i(t) \leq \mu_i^{(+)}, N_i(t) \geq N_i^{\mathrm{suf}}(T)]\big|\widehat{\mu}_i(t),\, N_i(t)\big]\Big] \\
&\leq \mathbb{E}\Big[\mathbb{E}\big[\mathbf{1}[\widehat{\mu}_i(t) \leq \mu_i^{(+)}, N_i(t) \geq N_i^{\mathrm{suf}}(T)] \\
&\qquad\qquad \mathbb{P}[\widetilde{\mu}_i(t) \geq \mu_L^{(-)}|\widehat{\mu}_i(t),\, N_i(t)]\big|\widehat{\mu}_i(t),\, N_i(t)\big]\Big] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\exp(-d(\mu_i^{(+)}, \mu_L^{(-)})N_i^{\mathrm{suf}}(T))\Big|\widehat{\mu}_i(t),\, N_i(t)\right]\right] \\
&\qquad\qquad\qquad\qquad \text{(by Lemma 26)} \\
&= \exp\left(-d(\mu_i^{(+)}, \mu_L^{(-)})N_i^{\mathrm{suf}}(T)\right) \\
&= T^{-1} \qquad\qquad \text{(by the definition of } N_i^{\mathrm{suf}}(T)\text{)}, \qquad\qquad (5.27)
\end{aligned}
$$

where we used the fact $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ for any random variables X and Y. Putting (5.26)–(5.27) together we obtain

$$
\mathbb{E}[(\mathrm{D})] \leq 1 + \frac{1}{d(\mu_i^{(+)}, \mu_i)} + \sum_{t=1}^{T} T^{-1},
$$

from which the inequality (5.16) follows. □

5.11.3 Proof of Lemma 22

It suffices to prove that for any $a, b > 0$

$$
\inf_{\epsilon_2 > 0}\left\{\frac{T^{-a\epsilon_2}}{b} + \epsilon_2\right\} = O\left(\frac{\log\log T}{\log T}\right).
$$

By letting $\epsilon_2 = (\log\log T)/(a\log T)$, we have

$$
\begin{aligned}
\inf_{\epsilon_2 > 0}\left\{\frac{T^{-a\epsilon_2}}{b} + \epsilon_2\right\} &= \inf_{\epsilon_2 > 0}\left\{\frac{e^{-a\epsilon_2 \log T}}{b} + \epsilon_2\right\} \\
&\leq \frac{e^{-\log\log T}}{b} + \frac{\log\log T}{a\log T} \\
&= \frac{1}{b\log T} + \frac{\log\log T}{a\log T} \\
&= O\left(\frac{\log\log T}{\log T}\right)
\end{aligned}
$$

and the proof is completed.

Chapter 6

Regret Lower Bound and Asymptotically Optimal Algorithm in Dueling Bandit Problem

In this chapter, we study the dueling bandit problem, a variation of the standard stochastic bandit problem where the feedback is limited to relative comparisons of a pair of arms. We introduce an asymptotic regret lower bound that is based on the information divergence. An algorithm that is inspired by the Deterministic Minimum Empirical Divergence algorithm is proposed, and its regret is analyzed. The proposed algorithm is found to be the first one with a regret upper bound that matches the lower bound. Experimental comparisons of dueling bandit algorithms show that the proposed algorithm significantly outperforms existing ones[*1]. The notation in this chapter is summarized in Table 6.1.

6.1 Motivation

In the multi-armed bandit problem, the availability of reward feedback from the selected arm is assumed. While it is desirable to obtain such a direct feedback from an arm, in some practical cases such direct feedback is not available. In this chapter, we consider a version of the standard stochastic bandit problem called the dueling bandit problem [Yue et al., 2009], in which the forecaster receives relative feedback, which specifies which of two arms is preferred. Although the dueling bandit problem was originally motivated by information retrieval applications, learning under relative feedback is universal to many fields, such as recommender systems [Gemmis et al., 2009], graphical design [Brochu et al., 2010], and natural language processing [Zaidan and Callison-Burch, 2011], which involve

[*1] The contents of this chapter were published in Komiyama et al. [2015a].

Table. 6.1. Notation used in Chapter 6.

$\mathbf{1}\{A\}$	:=	1 if A is true and 0 otherwise.
K	:=	Number of the arms.
$[K]$	:=	$\{1, 2, \ldots, K\}$.
T	:=	Number of the rounds.
$(l(t), m(t))$	:=	Pair of arms that is selected in round t.
$\widehat{X}_{l(t),m(t)}(t)$	:=	Feedback: which of $(l(t), m(t))$ is preferred.
$\mu_{i,j}$	:=	Probability that arm i is preferred to arm j.
$\Delta_{i,j}$	:=	$\mu_{i,j} - 1/2$.
$r(t)$	:=	$(\Delta_{1,l(t)} + \Delta_{1,m(t)})/2$.
$N_{i,j}(t)$	:=	Number of comparison between i and j: $\sum_{t'=1}^{t-1}(\mathbf{1}\{l(t') = i, m(t') = j\} + \mathbf{1}\{l(t') = j, m(t') = i\})$.
$\widehat{\mu}_{i,j}(t)$	:=	Empirical estimate of $\mu_{i,j}$: $(\sum_{t'=1}^{t-1}(\mathbf{1}\{l(t') = i, m(t') = j, \widehat{X}_{l(t'),m(t')}(t') = 1\} + \mathbf{1}\{l(t') = j, m(t') = i, \widehat{X}_{l(t'),m(t')}(t') = 0\}))/N_{i,j}(t)$.
$\mathcal{O}_i$	:=	$\{j \mid j \in [K], \mu_{i,j} < 1/2\}$.
$b^\star(i)$	:=	$\arg\min_{j \in \mathcal{O}_i} \frac{\Delta_{1,i} + \Delta_{1,j}}{d(\mu_{i,j}, 1/2)}$.
$\widehat{\mathcal{O}}_i(t)$	:=	$\{j \mid j \in [K] \setminus \{i\}, \widehat{\mu}_{i,j}(t) \leq 1/2\}$.
$\widehat{b}^\star(i)$	:=	Estimated $b^\star(i)$ (see Algorithms 10 and 12).
$d(p, q)$	:=	The KL divergence between Bernoulli distributions: $p \log(p/q) + (1-p) \log((1-p)/(1-q))$.
$d^+(p, q)$	:=	$d(p, q)$ if $p < q$ and 0 otherwise.
$\widehat{D}_i(t)$	:=	Empirical divergence: $\sum_{j \in \widehat{\mathcal{O}}_i(t)} N_{i,j}(t) d(\widehat{\mu}_{i,j}(t), 1/2)$.
$i^*(t)$	:=	$\arg\min_{i \in [K]} \widehat{D}_i(t)$.
$\widehat{D}^*(t)$	:=	$\widehat{D}_{i^*(t)}(t)$.
$f(K)$	:=	A non-negative function (see Algorithm 10).

explicit or implicit feedback provided by humans.

Related work: we briefly discuss the literature of the K-armed dueling bandit problem. The problem involves a preference matrix $M = \{\mu_{i,j}\} \in \mathbb{R}^{K \times K}$, whose ij entry $\mu_{i,j}$ corresponds to the probability that arm i is preferred to arm j.

Most algorithms assume that the preference matrix has certain properties. Interleaved Filter (IF) [Yue et al., 2012] and Beat the Mean Bandit (BTM) [Yue and Joachims, 2011], early algorithms proposed for solving the dueling bandit problem, require the arms to be totally ordered, that is, $i \succ j \Leftrightarrow \mu_{i,j} > 1/2$. Moreover, IF assumes *stochastic transitivity*: for any triple (i, j, k) with $i \succ j \succ k$, $\mu_{i,k} \geq \max\{\mu_{i,j}, \mu_{j,k}\}$. Unfortunately, stochastic transitivity does not hold in many real-world settings [Yue and Joachims, 2011]. BTM

relaxes this assumption by introducing *relaxed stochastic transitivity*: there exists $\gamma \geq 1$ such that for all pairs (j, k) with $1 \succ j \succ k$, $\gamma\mu_{1,k} \geq \max\{\mu_{1,j}, \mu_{j,k}\}$ holds. The drawback of BTM is that it requires the explicit value of γ on which the performance of the algorithm depends. Urvoy et al. [2013] considered a wide class of sequential learning problems with bandit feedback that includes the dueling bandit problem. They proposed the Sensitivity Analysis of VAriables for Generic Exploration (SAVAGE) algorithm, which empirically outperforms IF and BTM for moderate K. Among the several versions of SAVAGE, the one called Condorcet SAVAGE makes the *Condorcet assumption* and performed the best in their experiment. The Condorcet assumption is that there is a unique arm that is superior to the others. Unlike the two transitivity assumptions, the Condorcet assumption does not require the arms to be totally ordered and is less restrictive. IF, BTM, and SAVAGE either explicitly require the number of rounds T, or implicitly require T to determine the confidence level δ.

Recently, an algorithm called Relative Upper Confidence Bound (RUCB) [Zoghi et al., 2014b] was proven to have an $O(K \log T)$ regret bound under the Condorcet assumption. RUCB is based on the upper confidence bound index [Lai and Robbins, 1985, Agrawal, 1995b, Auer et al., 2002a] that is widely used in the field of bandit problems. RUCB is *horizonless*: it does not require T beforehand and runs for any duration. Zoghi et al. [2015] extended RUCB into the mergeRUCB algorithm under the Condorcet assumption as well as the assumption that a portion of the preference matrix is informative (i.e., different from 1/2). They reported that mergeRUCB outperformed RUCB when K was large. Ailon et al. [2014] proposed three algorithms named Doubler, MultiSBM, and Sparring. MultiSBM is endowed with an $O(K \log T)$ regret bound and Sparring was reported to outperform IF and BTM in their simulation. These algorithms assume that the pairwise feedback is generated from the non-observable utilities of the selected arms. The existence of the utility distributions associated with individual arms restricts the structure of the preference matrix.

In summary, most algorithms either has $O(K^2 \log T)$ regret under the Condorcet assumption (SAVAGE) or require additional assumptions to achieve $O(K \log T)$ regret (IF, BTM, MultiSBM, and mergeRUCB). To the best of our knowledge, RUCB is the only algorithm with an $O(K \log T)$ regret bound[*2]. The main difficulty of the dueling bandit problem lies in that, there are $K - 1$ candidates of actions to test "how good" each arm i is. A naive use of the confidence bound requires every pair of arms to be compared $O(\log T)$ times and yields an $O(K^2 \log T)$ regret bound.

Contribution: in this chapter, we propose an algorithm called Relative Minimum Empirical Divergence (RMED). The result here contributes to our understanding of the dueling bandit problem in the following three respects.

[*2] Zoghi et al. [2013] first proposed RUCB with an $O(K^2 \log T)$ regret bound and later modified it by adding a randomization procedure to assure $O(K \log T)$ regret in Zoghi et al. [2014b].

- **The asymptotical regret lower bound:** some studies (e.g., Yue et al. [2012]) have shown that the K-armed dueling bandit problem has a $\Omega(K \log T)$ regret lower bound. In this chapter, we further analyze this lower bound to obtain the optimal constant factor for models satisfying the Condorcet assumption. Furthermore, we show that the lower bound is the same under the total order assumption. This means that asymptotically optimal algorithms under the Condorcet assumption also achieve a lower bound of regret under the total order assumption even though such algorithms do not know that the arms are totally ordered.
- **An asymptotically optimal algorithm:** the regret of RMED is not only $O(K \log T)$, but also optimal in the sense that its constant factor matches the asymptotic lower bound under the Condorcet assumption. RMED is the first asymptotically optimal algorithm in the study of the dueling bandit problem.
- **Empirical performance assessment:** the performance of RMED is extensively evaluated by using five datasets: two synthetic datasets, one including preference data, and two including ranker evaluations in the information retrieval domain.

6.2 Problem Setup

The K-armed dueling bandit problem involves K arms that are indexed as $[K] = \{1, 2, \ldots, K\}$. Let $M \in \mathbb{R}^{K \times K}$ be a preference matrix whose ij entry $\mu_{i,j}$ corresponds to the probability that arm i is preferred to arm j. At each round $t = 1, 2, \ldots, T$, the forecaster selects a pair of arms $(l(t), m(t)) \in [K]^2$, then receives a relative feedback $\widehat{X}_{l(t),m(t)}(t) \sim \text{Bernoulli}(\mu_{l(t),m(t)})$ that indicates which of $(l(t), m(t))$ is preferred. By definition, $\mu_{i,j} = 1 - \mu_{j,i}$ holds for any $i, j \in [K]$ and $\mu_{i,i} = 1/2$.

Let $N_{i,j}(t)$ be the number of comparisons of pair (i, j) and $\widehat{\mu}_{i,j}(t)$ be the empirical estimate of $\mu_{i,j}$ at round t. In building statistics by using the feedback, we treat pairs without taking their order into consideration. Therefore, for $i \neq j$, $N_{i,j}(t) = \sum_{t'=1}^{t-1}(\mathbf{1}\{l(t') = i, m(t') = j\} + \mathbf{1}\{l(t') = j, m(t') = i\})$ and $\widehat{\mu}_{i,j}(t) = (\sum_{t'=1}^{t-1}(\mathbf{1}\{l(t') = i, m(t') = j, \widehat{X}_{l(t'),m(t')}(t') = 1\} + \mathbf{1}\{l(t') = j, m(t') = i, \widehat{X}_{l(t'),m(t')}(t') = 0\}))/N_{i,j}(t)$. For $j \neq i$, let $N_{i>j}(t)$ be the number of times i is preferred over j. Then, $\widehat{\mu}_{i,j}(t) = N_{i>j}(t)/N_{i,j}(t)$, where we set $0/0 = 1/2$ here. Let $\widehat{\mu}_{i,i}(t) = 1/2$.

Throughout this chapter, we will assume that the preference matrix has a Condorcet winner [Urvoy et al., 2013]. Here we call an arm i the Condorcet winner if $\mu_{i,j} > 1/2$ for any $j \in [K] \setminus \{i\}$. Without loss of generality, we will assume that arm 1 is the Condorcet winner. The set of preference matrices which have a Condorcet winner is denoted by $\mathcal{M}_{\text{C}}$. We also define the set of preference matrices satisfying the total order by $\mathcal{M}_{\text{o}} \subset \mathcal{M}_{\text{C}}$; that is, the relation $i \prec j \Leftrightarrow \mu_{i,j} < 1/2$ induces a total order iff $\{\mu_{i,j}\} \in \mathcal{M}_{\text{o}}$.

Let $\Delta_{i,j} = \mu_{i,j} - 1/2$. We define the regret per round as $r(t) = (\Delta_{1,i} + \Delta_{1,j})/2$ when the pair (i, j) is compared. The expectation of the cumulative regret, $\mathbb{E}[\text{Reg}(T)] =$

$\mathbb{E}\left[\sum_{t=1}^{T} r(t)\right]$ is used to measure the performance of an algorithm. The regret increases at each round unless the selected pair is $(l(t), m(t)) = (1, 1)$.

6.2.1 Regret lower bound in the dueling bandit problem

In this section we provide an asymptotic regret lower bound when $T \to \infty$. Let the superiors of arm i be a set $\mathcal{O}_i = \{j | j \in [K], \mu_{i,j} < 1/2\}$, that is, the set of arms that is preferred to i on average. The essence of the dueling bandit problem is how to eliminate each arm $i \in [K] \setminus \{1\}$ by making sure that arm i is not the Condorcet winner. To do so, the algorithm uses some of the arms in $\mathcal{O}_i$ and compares i with them.

A dueling bandit algorithm is strongly consistent for model $\mathcal{M} \subset \mathcal{M}_{\mathrm{C}}$ iff it has $\mathbb{E}[\mathrm{Reg}(T)] = o(T^a)$ regret for any $a > 0$ and any $M \in \mathcal{M}$. The following lemma is on the number of comparisons of suboptimal arm pairs.

Lemma 27. (The asymptotic lower bound on the number of suboptimal arm draws) *(i) Let an arm $i \in [K] \setminus \{1\}$ and preference matrix $M \in \mathcal{M}_{\mathrm{C}}$ be arbitrary. Given any strongly consistent algorithm for model $\mathcal{M}_{\mathrm{C}}$, we have*

$$\mathbb{E}\left\{\sum_{j \in \mathcal{O}_i} d(\mu_{i,j}, 1/2) N_{i,j}(T)\right\} \geq (1 - o(1)) \log T, \tag{6.1}$$

where $d(p, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ is the KL divergence between two Bernoulli distributions with parameters p and q. (ii) Furthermore, inequality (6.1) *holds for any $M \in \mathcal{M}_{\mathrm{o}}$ given any strongly consistent algorithm for $\mathcal{M}_{\mathrm{o}}$.*

Lemma 27 states that, for arbitrary arm $j \in \mathcal{O}_i$, an algorithm needs to make $\log T / d(\mu_{i,j}, 1/2)$ comparisons between arms i and j to be convinced that arm i is inferior to arm j and thus i is not the Condorcet winner. Since the regret increase per round of comparing arm i with j is $(\Delta_{1,i} + \Delta_{1,j})/2$, eliminating arm i by comparing it with j incurs a regret of

$$\frac{(\Delta_{1,i} + \Delta_{1,j}) \log T}{2 d(\mu_{i,j}, 1/2)}. \tag{6.2}$$

Therefore, the total regret is bounded from below by comparing each arm i with an arm j that minimizes (6.2), and the regret lower bound is formalized in the following theorem.

Theorem 28. (The asymptotic regret lower bound) *(i) Let the preference matrix $M \in \mathcal{M}_{\mathrm{C}}$ be arbitrary. For any strongly consistent algorithm for model $\mathcal{M}_{\mathrm{C}}$,*

$$\liminf_{T \to \infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \geq \sum_{i \in [K] \setminus \{1\}} \min_{j \in \mathcal{O}_i} \frac{\Delta_{1,i} + \Delta_{1,j}}{2 d(\mu_{i,j}, 1/2)} \tag{6.3}$$

holds. (ii) Furthermore, inequality (6.3) *holds for any $M \in \mathcal{M}_{\mathrm{o}}$ given any strongly consistent algorithm for $\mathcal{M}_{\mathrm{o}}$.*

The proof of Lemma 27 and Theorem 28 can be found in Section 6.8. The proof of Lemma 27 is similar to that of Lai and Robbins [1985] (c.f., Section 2.3.2) for the standard multi-armed bandit problem but differs in the following point that is characteristic to the dueling bandit. To achieve a small regret in the dueling bandit, it is necessary to compare the arm i with itself if i is the Condorcet winner. However, we trivially know that $\mu_{i,i} = 1/2$ without sampling and such a comparison yields no information to distinguish possible preference matrices. We can avoid this difficulty by evaluating $N_{i,j}$ and $N_{i,i}$ in different ways.

6.3 RMED1 Algorithm

In this section, we first introduce the notion of empirical divergence. Then, on the basis of the empirical divergence, we formulate the RMED1 algorithm.

6.3.1 Empirical divergence and likelihood function

In inequality (6.1) of Section 6.2.1, we have seen that $\sum_{j \in \mathcal{O}_i} d(\mu_{i,j}, 1/2) N_{i,j}(T)$, the sum of the divergence between $\mu_{i,j}$ and $1/2$ multiplied by the number of comparisons between i and j, is the characteristic value that defines the minimum number of comparisons. The empirical estimate of this value is fundamentally useful for evaluating how unlikely arm i is to be the Condorcet winner. Let the opponents of arm i at round t be the set $\widehat{\mathcal{O}}_i(t) = \{j | j \in [K] \setminus \{i\}, \widehat{\mu}_{i,j}(t) \leq 1/2\}$. Note that, unlike the superiors $\mathcal{O}_i$, the opponents $\widehat{\mathcal{O}}_i(t)$ for each arm i are defined in terms of the empirical averages, and thus the algorithms know who the opponents are. Let the empirical divergence be

$$\widehat{D}_i(t) = \sum_{j \in \widehat{\mathcal{O}}_i(t)} N_{i,j}(t) d(\widehat{\mu}_{i,j}(t), 1/2).$$

The value $\exp\left(-\widehat{D}_i(t)\right)$ can be considered as the "likelihood" that arm i is the Condorcet winner. Let $i^*(t) = \arg\min_{i \in [K]} \widehat{D}_i(t)$ (ties are broken arbitrarily) and $\widehat{D}^*(t) = \widehat{D}_{i^*(t)}(t)$. By definition, $\widehat{D}^*(t) \geq 0$. RMED is inspired by the Deterministic Minimum Empirical Divergence (DMED) algorithm [Honda and Takemura, 2010]. DMED, which is designed for solving the standard multi-armed bandit problem, draws arms that may be the best one with probability $\Omega(1/t)$, whereas RMED in the dueling bandit problem draws arms that are likely to be the Condorcet winner with probability $\Omega(1/t)$. Namely, any arm i that satisfies

$$\mathcal{J}_i(t) = \{\widehat{D}_i(t) - \widehat{D}^*(t) \leq \log t + f(K)\}$$

is the candidate of the Condorcet winner and will be drawn soon. Here, $f(K)$ can be any non-negative function of K that is independent of t. The effect of $f(K)$ is discussed in Section 6.7. Algorithm 10 lists the main routine of RMED. There are several versions of

Algorithm 10 Relative Minimum Empirical Divergence (RMED) Algorithm

Input: K arms, $f(K) \geq 0$. $\alpha > 0$ (RMED2FH, RMED2). T (RMED2FH).
$L \leftarrow \begin{cases} 1 & \text{(RMED1, RMED2)} \\ \lceil \alpha \log \log T \rceil & \text{(RMED2FH)} \end{cases}$.
Initial phase: draw each pair of arms L times. At the end of this phase, $t = L(K-1)K/2$.
if RMED2FH **then**
  For each arm $i \in [K]$, fix $\widehat{b}^\star(i)$ by (6.5).
end if
$L_C, L_R \leftarrow [K], L_N \leftarrow \emptyset$.
while $t \leq T$ **do**
  if RMED2 **then**
    Draw all pairs (i, j) until it reaches $N_{i,j}(t) \geq \alpha \log \log t$. $t \leftarrow t + 1$ for each draw.
  end if
  for $l(t) \in L_C$ in an arbitrarily fixed order **do**
    Select $m(t)$ by using $\begin{cases} \text{Algorithm 11} & \text{(RMED1)} \\ \text{Algorithm 12} & \text{(RMED2, RMED2FH)} \end{cases}$.
    Draw arm pair $(l(t), m(t))$.
    $L_R \leftarrow L_R \setminus \{l(t)\}$.
    $L_N \leftarrow L_N \cup \{j\}$ (without a duplicate) for any $j \notin L_R$ such that $\mathcal{J}_j(t)$ holds.
    $t \leftarrow t + 1$.
  end for
  $L_C, L_R \leftarrow L_N$, $L_N \leftarrow \emptyset$.
end while

RMED. First, we introduce RMED1. RMED1 initially compares all pairs once (initial phase). Let $T_{\text{init}} = (K-1)K/2$ be the last round of the initial phase. From $t = T_{\text{init}} + 1$, it selects the arm by using a loop. $L_C = L_C(t)$ is the set of arms in the current loop, and $L_R = L_R(t) \subset L_C(t)$ is the remaining arms of L_C that have not been drawn yet in the current loop. $L_N = L_N(t)$ is the set of arms that are going to be drawn in the next loop. An arm i is put into L_N when it satisfies $\{\mathcal{J}_i(t) \cap \{i \notin L_R(t)\}\}$. By definition, at least one arm (i.e. $i^*(t)$ at the end of the current loop) is put into L_N in each loop. For arm $l(t)$ in the current loop, RMED1 selects $m(t)$ (i.e. the comparison target of $l(t)$) determined by Algorithm 11.

The following theorem, which is proven in Section 6.5, describes a regret bound of RMED1.

Algorithm 11 RMED1 subroutine for selecting $m(t)$

$\widehat{\mathcal{O}}_{l(t)}(t) \leftarrow \{j \in [K] \setminus \{l(t)\} | \widehat{\mu}_{l(t),j}(t) \leq 1/2\}$
if $i^*(t) \in \widehat{\mathcal{O}}_{l(t)}(t)$ or $\widehat{\mathcal{O}}_{l(t)}(t) = \emptyset$ **then**
  $m(t) \leftarrow i^*(t)$.
else
  $m(t) \leftarrow \arg\min_{j \neq l(t)} \widehat{\mu}_{l(t),j}(t)$.
end if

Theorem 29. *For any sufficiently small $\delta > 0$, the regret of RMED1 is bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \in [K] \setminus \{1\}} \frac{((1+\delta)\log T + f(K))\Delta_{1,i}}{2d(\mu_{i,1}, 1/2)} + O\left(\frac{K}{\delta^2}\right),$$

when we view model parameters $\{\mu_{i,j}\}_{i,j \in [K]}$ and K as constants that are independent of T. Therefore, by letting $\delta = \log^{-1/3} T$, we obtain

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \in [K] \setminus \{1\}} \frac{\Delta_{1,i} \log T}{2d(\mu_{i,1}, 1/2)} + O(K \log^{2/3} T).$$

as a function of T.

6.3.2 Gap between the constant factor of RMED1 and the lower bound

From the lower bound of Theorem 28, the $O(K \log T)$ regret bound of RMED1 is optimal up to a constant factor. Moreover, the constant factor matches the regret lower bound of Theorem 28 if $b^\star(i) = 1$ for all $i \in [K] \setminus \{1\}$ where

$$b^\star(i) = \arg\min_{j \in \mathcal{O}_i} \frac{\Delta_{1,i} + \Delta_{1,j}}{d(\mu_{i,j}, 1/2)}. \tag{6.4}$$

Here we define $d^+(p,q) = d(p,q)$ if $p < q$ and 0 otherwise, and $x/0 = +\infty$. Note that, there can be ties that minimize the RHS of (6.4). In that case, we may choose any of the ties as $b^\star(i)$ to eliminate arm i. For ease of explanation, we henceforth will assume that $b^\star(i)$ is unique, but our results can be easily extended to the case of ties.

We claim that $b^\star(i) = 1$ holds in many cases for the following mathematical and practical reasons. (i) The regret of drawing a pair (i,j), $j \neq 1$, is $(\Delta_{1,i}+\Delta_{1,j})/2$, whereas it is simply $\Delta_{1,i}/2$ for the pair $(i,1)$. Thus, $d^+(\mu_{i,j}, 1/2)$ has to be much larger than $d^+(\mu_{i,1}, 1/2)$ in order to satisfy $b^\star(i) = j$. (ii) The Condorcet winner usually wins over the other arms by a large margin, and therefore, $d^+(\mu_{i,1}, 1/2) \geq d^+(\mu_{i,j}, 1/2)$. For example, in the preference matrix of Example 1 (Table 6.2(a)), $b^\star(3) = 1$ as long as $q < 0.79$. Example 2 (Table 6.2(b)) is a preference matrix based on six retrieval functions in the full-text search engine

of ArXiv.org [Yue and Joachims, 2011][*3]. In Example 2, $b^\star(i) = 1$ holds for all i, even though $\mu_{1,4} < \mu_{2,4}$. In the case of a 16-ranker evaluation based on the Microsoft Learning to Rank dataset (details are given in Section 6.4), occasionally $b^\star(i) \neq 1$ occurs, but the difference between the regrets of drawing arm 1 and $b^\star(i)$ is fairly small (smaller than 1.2% on average). Nevertheless, there are some cases in which comparing arm i with 1 is not a clever idea. Example 3 (Table 6.2(c)) is a toy example in which comparing arm i with $b^\star(i) \neq 1$ makes a large difference. In Example 3, it is clearly better to draw pairs (2, 4), (3, 2) and (4, 3) to eliminate arms 2, 3, and 4, respectively. Accordingly, it is still interesting to consider an algorithm that reduces regret by comparing arm i with $b^\star(i)$.

Table. 6.2. Three preference matrices. In each example, the value at row i, column j is $\mu_{i,j}$.

	1	2	3
1	0.5	0.7	0.7
2	0.3	0.5	q
3	0.3	1-q	0.5

(a) Example 1

	1	2	3	4	5	6
1	0.50	0.55	0.55	0.54	0.61	0.61
2	0.45	0.50	0.55	0.55	0.58	0.60
3	0.45	0.45	0.50	0.54	0.51	0.56
4	0.46	0.45	0.46	0.50	0.54	0.50
5	0.39	0.42	0.49	0.46	0.50	0.51
6	0.39	0.40	0.44	0.50	0.49	0.50

(b) Example 2

	1	2	3	4
1	0.5	0.6	0.6	0.6
2	0.4	0.5	0.9	0.1
3	0.4	0.1	0.5	0.9
4	0.4	0.9	0.1	0.5

(c) Example 3

6.3.3 RMED2 Algorithm

We here propose RMED2, which gracefully estimates $b^\star(i)$ during a bandit game and compares arm i with $b^\star(i)$. RMED2 and RMED1 share the main routine (Algorithm 10). The subroutine of RMED2 for selecting $m(t)$ is shown in Algorithm 12. Unlike RMED1, RMED2 keeps drawing pairs of arms (i, j) at least $\alpha \log\log t$ times (Line 10 in Algorithm

[*3] In the original preference matrix of Yue and Joachims [2011], $\mu_{2,4} \neq 1 - \mu_{4,2}$. To satisfy $\mu_{2,4} = 1 - \mu_{4,2}$, we replaced $\mu_{2,4}$ and $\mu_{4,2}$ of the original with $(\mu_{2,4} - \mu_{4,2} + 1)/2$ and $(\mu_{4,2} - \mu_{2,4} + 1)/2$, respectively.

Algorithm 12 Subroutine for selecting $m(t)$ in RMED2 and RMED2FH

if RMED2 **then**
  Update $\widehat{b}^\star(l(t))$ by (6.5).
end if
$\widehat{\mathcal{O}}_{l(t)}(t) \leftarrow \{j \in [K] \setminus \{l(t)\} | \widehat{\mu}_{l(t),j}(t) \leq 1/2\}$.
if $\widehat{b}^\star(l(t)) \in \widehat{\mathcal{O}}_{l(t)}(t)$ and $\begin{cases} N_{l(t),i^*(t)}(t) \geq N_{l(t),\widehat{b}^\star(l(t))}(t)/\log\log t & \text{(RMED2)} \\ N_{l(t),i^*(t)}(t) \geq N_{l(t),\widehat{b}^\star(l(t))}(t)/\log\log T & \text{(RMED2FH)} \end{cases}$ **then**
  $m(t) \leftarrow \widehat{b}^\star(l(t))$.
else
  Select $m(t)$ by using Algorithm 11.
end if

10). The regret of this exploration is insignificant since $O(\log\log T) = o(\log T)$. Once all pairs have been explored more than $\alpha \log\log t$ times, RMED2 goes to the main loop. RMED2 determines $m(t)$ by using Algorithm 12 based on the estimate of $b^\star(i)$ given by

$$\widehat{b}^\star(i) = \arg\min_{j \in [K]\setminus\{i\}} \frac{\widehat{\Delta}_{i^*(t),i} + \widehat{\Delta}_{i^*(t),j}}{d^+(\widehat{\mu}_{i,j}(t), 1/2)}, \tag{6.5}$$

where ties are broken arbitrarily, $\widehat{\Delta}_{i,j} = 1/2 - \widehat{\mu}_{i,j}$ and we set $x/0 = +\infty$. Intuitively, RMED2 tries to select $m(t) = \widehat{b}^\star(i)$ for most rounds, and occasionally explores $i^*(t)$ in order to reduce the regret increase when RMED2 fails to estimate the true $b^\star(i)$ correctly.

6.3.4 RMED2FH algorithm

Although we believe that the regret of RMED2 is asymptotically optimal, the analysis of RMED2 is a little bit complicated since it sometimes breaks the main loop and explores from time to time. For ease of analysis, we here propose RMED2 Fixed Horizon (RMED2FH, Algorithm 10 and 12), which is a "static" version of RMED2. Essentially, RMED2 and RMED2FH have the same mechanism. The differences are that (i) RMED2FH conducts an $\alpha \log\log T$ exploration in the initial phase. After the initial phase (ii) $\widehat{b}^\star(i)$ for each i is fixed throughout the game. Note that, unlike RMED1 and RMED2, RMED2FH requires the number of rounds T beforehand to conduct the initial $\alpha \log\log T$ draws of each pair. The following Theorem shows the regret of RMED2FH that matches the lower bound of Theorem 28.

Theorem 30. *For any sufficiently small $\delta > 0$, the regret of RMED2FH is bounded as*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \in [K] \setminus \{1\}} \frac{(\Delta_{1,i} + \Delta_{1,b^\star(i)})((1+\delta)\log T)}{2d(\mu_{i,b^\star(i)}, 1/2)} + O(\alpha K^2 \log\log T) + O\left(\frac{K \log T}{\log\log T}\right) + O\left(\frac{K}{\delta^2}\right),$$

when we view model parameters $\{\mu_{i,j}\}_{i,j\in[K]}$ and K as constants that are independent of T.

By setting $\delta = O((\log T)^{-1/3})$ we obtain

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i \in [K] \setminus \{1\}} \frac{(\Delta_{1,i} + \Delta_{1,b^\star(i)})\log T}{2d(\mu_{i,b^\star(i)}, 1/2)} + O(\alpha K^2 \log\log T) + O\left(\frac{K \log T}{\log\log T}\right). \quad (6.6)$$

Note that, the last two terms in the RHS of (6.6) are $o(\log T)$. From Theorems 28 and 30 we see that (i) RMED2FH is asymptotically optimal under the Condorcet assumption and (ii) the logarithmic term on the regret bound of RMED2FH cannot be improved even if the arms are totally ordered and the forecaster knows of the existence of the total order. The proof sketch of Theorem 30 is in Section 6.5.

6.4 Experimental Evaluation

To evaluate the empirical performance of RMED, we conducted simulations[*4] with five bandit datasets (preference matrices). The datasets are as follows:

Six rankers is the preference matrix based on the six retrieval functions in the full-text search engine of ArXiv.org (Table 6.2(b)).

Cyclic is the artificial preference matrix shown in Table 6.2(c). This matrix is designed so that the comparison of i with 1 is not optimal.

Arithmetic dataset involves eight arms with $\mu_{i,j} = 0.5 + 0.05(j-i)$ and has a total order.

Sushi dataset is based on the Sushi preference dataset [Kamishima, 2003] that contains the preferences of $5,000$ Japanese users as regards 100 types of sushi. We extracted the 16 most popular types of sushi and converted them into arms with $\mu_{i,j}$ corresponding to the ratio of users who prefer sushi i over j. The Condorcet winner is the mildly-fatty tuna (chu-toro).

MSLR: we tested submatrices of a 136×136 preference matrix from Zoghi et al. [2015], which is derived from the Microsoft Learning to Rank (MSLR) dataset [Microsoft Research, 2010, Qin et al., 2010] that consists of relevance information between queries and documents with more than 30K queries. Zoghi et al. [2015] created a finite set of rankers, each of which corresponds to a ranking feature in the base dataset. The value $\mu_{i,j}$ is

[*4] The source code of the simulations is available at https://github.com/jkomiyama/duelingbanditlib.

the probability that the ranker i beats ranker j based on the navigational click model [Hofmann et al., 2013]. We randomly extracted $K = 16, 64$ rankers in our experiments and made sub preference matrices. The probability that the Condorcet winner exists in the subset of the rankers is high (more than 90%, c.f. Figure 1 in Zoghi et al. [2014a]), and we excluded the relatively small case where the Condorcet winner does not exist.

A Condorcet winner exists in all datasets. In the experiments, the regrets of the algorithms were averaged over $1,000$ runs (Six rankers, Cyclic, Arithmetic, and Sushi), or 100 runs (MSLR).

6.4.1 Comparison among algorithms

We compared the IF, BTM with $\gamma = 1.2$, RUCB with $\alpha = 0.51$, Condorcet SAVAGE with $\delta = 1/T$, MultiSBM and Sparring with $\alpha = 3$, and RMED algorithms. We set $f(K) = 0.3K^{1.01}$ for all RMED algorithms and set $\alpha = 3$ for RMED2 and RMED2FH. The effect of $f(K)$ is studied in Section 6.7. Note that IF and BTM assume a total order among arms, which is not the case with the Cyclic, Sushi, and MSLR datasets. MultiSBM and Sparring assume the existence of the utility of each arm, which does not allow a cyclic preference that appears in the Cyclic dataset.

Figure 6.1 plots the regrets of the algorithms. In all datasets RMED significantly outperforms RUCB, the next best excluding the different versions of RMED. Notice that the plots are on a base 10 log-log scale. In particular, regret of RMED1 is more than twice smaller than RUCB on all datasets other than Cyclic, in which RMED2 performs much better. Among the RMED algorithms, RMED1 outperforms RMED2 and RMED2FH on all datasets except for Cyclic, in which comparing arm $i \neq 1$ with arm 1 is inefficient. RMED2 outperforms RMED2FH in the five of six datasets: this could be due to the fact that RMED2FH does not update $\widehat{b}^\star(i)$ for ease of analysis.

6.4.2 RMED and asymptotic bound

Figure 6.2 compares the regret of RMED with two asymptotic bounds. LB1 denotes the regret bound of RMED1. TrueLB is the asymptotic regret lower bound given by Theorem 28.

RMED1 and RMED2: when $T \to \infty$, the slope of RMED1 should converge to LB1, and the ones of RMED2 and RMED2FH should converge to TrueLB. On Six rankers, LB1 is exactly the same as TrueLB, and the slope of RMED1 converges to this TrueLB. In Cyclic, the slope of RMED2 converges to TrueLB, whereas that of RMED1 converges to LB1, from which we see that RMED2 is actually able to estimate $b^\star(i) \neq 1$ correctly. In MSLR $K = 16$, LB1 and TrueLB are very close (the difference is less than 1.2%), and RMED1 and RMED2 converge to these lower bounds.

RMED2FH with different values of α: we also tested RMED2FH with several values

of α. On the one hand, with $\alpha = 1$, the initial phase of RMED2FH is too short to identify $b^\star(i)$; as a result it performs poorly on the Cyclic dataset. On the other hand, with $\alpha = 10$, the initial phase was too long, which incurs a practically non-negligible regret on the MSLR $K = 16$ dataset. We also tested several values of parameter α in RMED2FH. We omit plots of RMED2 with $\alpha = 1$, 10 for the sake of readability, but we note that in our datasets the performance of RMED2 is always better than or comparable with the one of RMED2FH under the same choice of α, although the optimality of RMED2 is not proved unlike RMED2FH.

6.5 Regret Analysis

This section provides two lemmas essential for the regret analysis of RMED algorithms and proves the asymptotic optimality of RMED1 based on these lemmas. A proof sketch on the optimal regret of RMED2FH is also given.

The crucial property of RMED is that, by constantly comparing arms with the opponents, the true Condorcet winner (arm 1) actually beats all the other arms with high probability. Let

$$\mathcal{U}(t) = \bigcap_{i \in [K] \setminus \{1\}} \{\widehat{\mu}_{1,i}(t) > 1/2\}.$$

Under $\mathcal{U}(t)$, $\widehat{\mu}_{i,1}(t) = 1 - \widehat{\mu}_{1,i}(t) < 1/2$ for all $i \in [K] \setminus \{1\}$, and thus, $\widehat{D}_i(t) > 0$. Therefore, $\mathcal{U}(t)$ implies that $i^*(t) = \arg\min_{i \in [K]} \widehat{D}_i(t)$ is unique with $i^*(t) = 1$ and $\widehat{D}^*(t) = \widehat{D}_1(t) = 0$. Lemma 31 below shows that the average number of rounds that $\mathcal{U}^c(t)$ occurs is constant in T, where the superscript c denotes the complement.

Lemma 31. *When RMED1 or RMED2FH is run, the following inequality holds:*

$$\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}\{\mathcal{U}^c(t)\}\right] = O(\mathrm{e}^{AK}) = O(1),$$

where $A = A(\{\mu_{i,j}\}) > 0$ is a constant as a function of T.

Note that, since RMED2FH draws each pair $\lceil \alpha \log\log T \rceil$ times in the initial phase, we define $T_{\mathrm{init}} = \lceil \alpha \log\log T \rceil (K-1)K/2$ for RMED2FH. We give a proof of this lemma in Section 6.9. Intuitively, this lemma can be proved from the facts that arm 1 is drawn within roughly $\mathrm{e}^{\widehat{D}_1(t)}$ rounds and $\widehat{D}_1(t)$ is not very large with high probability.

Next, for $i \in [K] \setminus \{1\}$ and $j \in \mathcal{O}_i$, let

$$N_{i,j}^{\mathrm{Suf}}(\delta) = \frac{(1+\delta)\log T + f(K)}{d(\mu_{i,j}, 1/2)} + 1,$$

which is a sufficient number of comparisons of i with j to be convinced that the arm i is not the Condorcet winner. The following lemma states that if pair (i,j) is drawn $N_{i,j}^{\mathrm{Suf}}(\delta)$ times then i is rarely selected as $l(t)$ again.

Lemma 32. *When RMED1 or RMED2FH is run, for $i \in [K] \setminus \{1\}$, $j \in \mathcal{O}_i$,*

$$\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}\{l(t) = i, N_{i,j}(t) \geq N_{i,j}^{\mathrm{Suf}}(\delta)\}\right] = O\left(\frac{1}{\delta^2}\right).$$

We prove this lemma in Section 6.10 based on the Chernoff bound.

Now we can derive the regret bound of RMED1 based on these lemmas.

Proof of Theorem 29: since $\mathcal{U}(t)$ implies $m(t) = 1$ in RMED1, the regret increase per round can be decomposed as

$$r(t) = \mathbf{1}\{\mathcal{U}^c(t)\} + \sum_{i \in [K]\setminus\{1\}} \frac{\Delta_{1,i}}{2} \mathbf{1}\{l(t) = i, m(t) = 1, \mathcal{U}(t)\}.$$

Using Lemmas 31 and 32, we obtain

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &\leq T_{\mathrm{init}} + \sum_{t=T_{\mathrm{init}}+1}^{T} [r(t)] \\
&\leq \frac{K(K-1)}{2} + \mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}\{\mathcal{U}^c(t)\}\right] \\
&\quad + \sum_{i \in [K]\setminus\{1\}} \frac{\Delta_{1,i}}{2} \left(N_{i,1}^{\mathrm{Suf}}(\delta) + \sum_{t=1}^{T} \mathbf{1}[l(t) = i, m(t) = 1, N_{i,1}(t) \geq N_{i,1}^{\mathrm{Suf}}(\delta)]\right) \\
&\leq \frac{K(K-1)}{2} + O(1) + \sum_{i \in [K]\setminus\{1\}} \frac{\Delta_{1,i}}{2} \left(N_{i,1}^{\mathrm{Suf}}(\delta) + O\left(\frac{1}{\delta^2}\right) + K\right),
\end{aligned}$$

which immediately completes the proof of Theorem 29. □

We also prove Theorem 30 on the optimality of RMED2FH based on Lemmas 31 and 32. Because the full proof in Section 6.11 is a little bit lengthy, here we give its brief sketch.

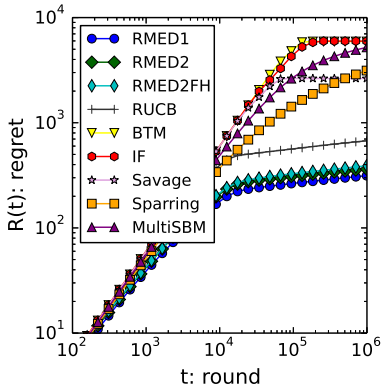
Proof sketch of Theorem 30 (RMED2FH): similar to Theorem 29, we use the fact that the $\mathcal{U}^c(t)$ does not occur very often (i.e., Lemma 31). Under $\mathcal{U}(t)$, we decompose the regret into the contributions of each arm $i \in [K] \setminus \{1\}$. There exists $C_2 > 0$ such that, for each $l(t) = i$, (i) with probability $1 - O((\log T)^{-C_2})$ RMED2FH successfully estimates $\widehat{b^\star}(i) = b^\star(i)$ and selects $m(t) = b^\star(i)$ for most rounds. The optimal $O(\log T)$ term comes from the comparison of i and $b^\star(i)$. Arm 1 is also drawn for $O(\log T / \log\log T) = o(\log T)$ times. On the other hand, (ii) with probability $O((\log T)^{-C_2})$, RMED2FH fails to estimate $b^\star(i)$ correctly. By occasionally comparing arm i with arm 1, we can bound the regret increase by $O(\log T \log\log T)$. Since $O((\log T)^{-C_2} \times \log T \log\log T) = o(\log T)$, this regret does not affect the $O(\log T)$ factor.

6.6 Discussion

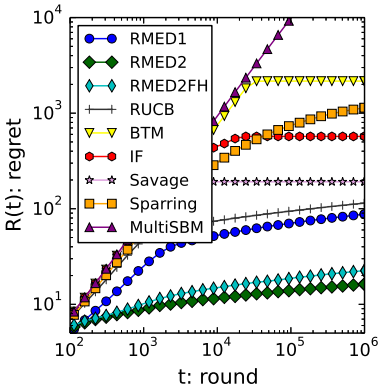
We proved the asymptotic regret lower bound in the dueling bandit problem. The RMED algorithm is based on the likelihood that the arm is the Condorcet winner. RMED is proven to have a matching regret upper bound. The empirical evaluation revealed that RMED significantly outperforms the state-of-the-art algorithms. To conclude this chapter, we mention the following directions of future work.

First, when a Condorcet winner does not necessarily exist, the Copeland bandits [Urvoy et al., 2013] are a natural extension of our problem. Thus, seeking an effective algorithm for solving this problem will be interesting. As is well known in the field of voting theory, there are several other criteria of winners that are incompatible with the Condorcet / Copeland bandits, such as the Borda winner [Urvoy et al., 2013]. Comparing several criteria or developing an algorithm that outputs more than one of these winners should be interesting directions of future work.

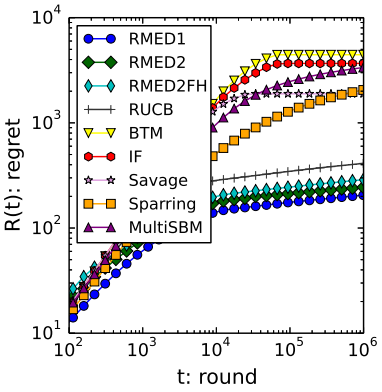
Second, another direction is sequential preference elicitation problems under relative feedback that goes beyond the binary preference over pairs, such as multiscale feedback and/or preferences among three or more items.



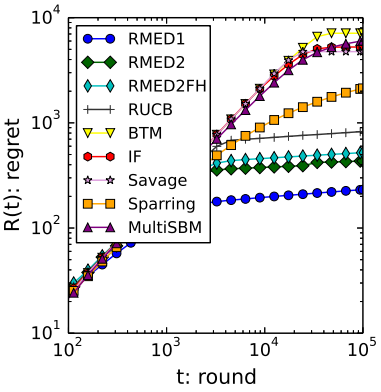
(a) Six rankers



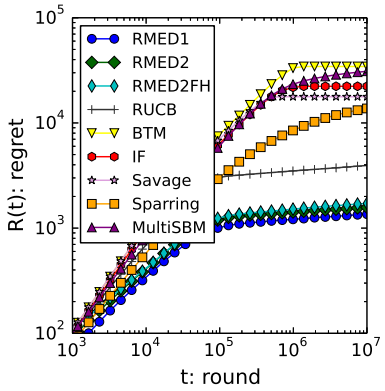
(b) Cyclic



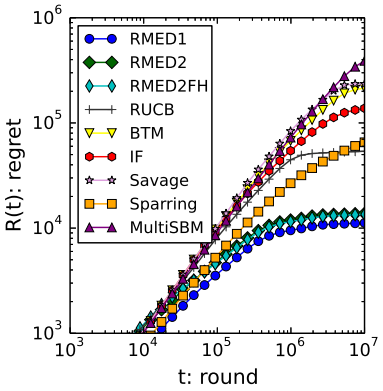
(c) Arithmetic



(d) Sushi

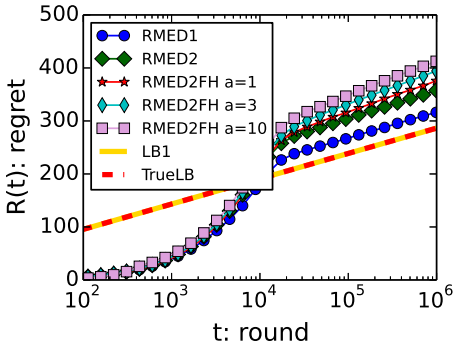


(e) MSLR $K = 16$

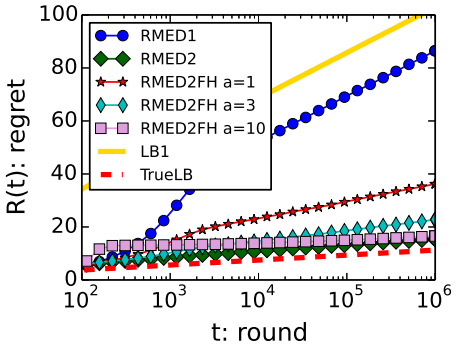


(f) MSLR $K = 64$

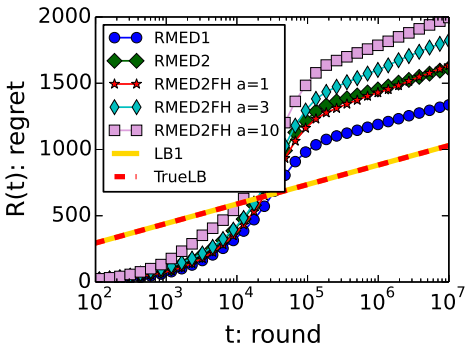
Fig. 6.1. Regret-round log-log plots of algorithms.



(a) Six rankers



(b) Cyclic



(c) MSLR $K = 16$

Fig. 6.2. Regret-round semilog plots of RMED compared with theoretical bounds. We set $f(K) = 0.3K^{1.01}$ for all algorithms, and $\alpha = 3$ for RMED2.

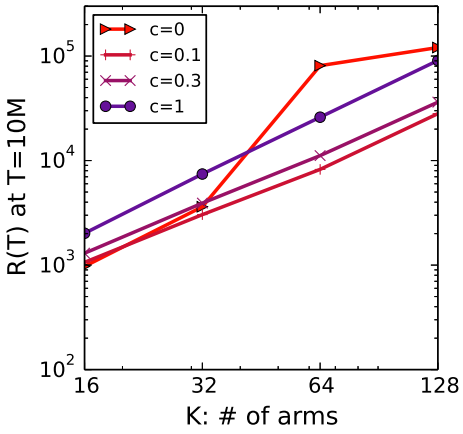


Fig. 6.3. Performance of RMED1 algorithm with several values of c. The plot shows the regret at $T = 10^7$ in the MSLR dataset with $K = 16, 32, 64$, and 128.

6.7 Experiment: Dependence on $f(K)$

The event $\mathcal{U}^c(t)$ implies a failure in identifying the Condorcet winner (i.e., $1 \neq i^*(t)$). Although $\mathbb{E}[\sum_{t=1}^{T} \mathcal{U}^c(t)] = O(\mathrm{e}^{AK})$ is a constant function of T for any non-negative $f(K)$ (see Lemma 31), this term is not negligible with large K. The introduction of $f(K) = K^{1+\epsilon}$ with $\epsilon > 0$ can remedy this problem. Although we cannot prove, intuitively the term can be exponentially small as $O(\mathrm{e}^{AK-f(K)})$ as the following argument. The exponential term is related to how long does it takes to escape from the underestimation of arm 1, which is exponential to the empirical divergence (Inequality (6.13)). Since $\log t - f(K) = \log(t/e^{f(K)})$, the t in (6.13) can be replaces by $t/e^{f(K)}$, which implies an exponentially fast escape from the underestimated state. To practically evaluate the effect of $f(K)$, we set $f(K) = cK^{1.01}$ and studied several values of c with the MSLR dataset (Figure 6.3). In the case of $c = 0$, the regret for $K = 128$ becomes 100 times that for $K = 16$, which implies that the exponential dependence $O(\mathrm{e}^{AK})$ may not be an artifact of the proof. On the other hand, the results for $c = 0.1$, 0.3, and 1 indicate that this term can be much improved by simply letting c be a small positive value.

6.8 Proofs on Regret Lower Bound

6.8.1 Proof of Lemma 27

Proof. Proof of Lemma 27

Let $i \in [K] \setminus \{1\}$ be arbitrary and $M = \{\mu_{i,j}\}$ be an arbitrary preference matrix. We

consider a modified preference matrix M' in which the probabilities related to arm i are different from M. Let $\mathcal{O}'_i = \{j | j \in [K], \mu_{i,j} \leq 1/2\}$, that is, $\mathcal{O}'_i = \mathcal{O}_i \cup \{j | j \in [K], \mu_{i,j} = 1/2\}$. For $j \in \mathcal{O}'_i$, ij element of M' is $\mu'_{i,j}$ such that

$$d^+(\mu_{i,j}, \mu'_{i,j}) = d(\mu_{i,j}, 1/2) + \epsilon. \tag{6.7}$$

Such a $\mu'_{i,j} > 1/2$ uniquely exists for sufficiently small $\epsilon > 0$ by the monotonicity and continuity of the KL divergence. For $j \notin \mathcal{O}'_i$, let $\mu'_{i,j} = \mu_{i,j}$. Note that, unlike the original bandit problem, in the modified bandit problem the Condorcet winner is not arm 1 but arm i. Moreover, if $M \in \mathcal{M}_{\mathrm{o}}$ then $M' \in \mathcal{M}_{\mathrm{o}}$.

Notation: now, let $\widehat{X}^m_{i,j} \in \{0, 1\}$ be the result of m-th draw of the pair (i, j),

$$\widehat{\mathrm{KL}}_j(n_j) = \sum_{m=1}^{n_j} \log \left(\frac{\widehat{X}^m_{i,j}\mu_{i,j} + (1 - \widehat{X}^m_{i,j})(1 - \mu_{i,j})}{\widehat{X}^m_{i,j}\mu'_{i,j} + (1 - \widehat{X}^m_{i,j})(1 - \mu'_{i,j})} \right),$$

and $\widehat{\mathrm{KL}}(\{n_j\}_{j \in \mathcal{O}'_i}) = \sum_{j \in \mathcal{O}'_i} \widehat{\mathrm{KL}}_j(n_j)$, and $\mathbb{P}'$, $\mathbb{E}'$ be the probability and the expectation with respect to the modified bandit game. Let us define the events

$$\begin{aligned}
\mathcal{D}_1 &= \left\{ \sum_{j \in \mathcal{O}'_i} N_{i,j}(T) d(\mu_{i,j}, \mu'_{i,j}) < (1 - \epsilon) \log T, N_{i,i}(T) < \sqrt{T} \right\}, \\
\mathcal{D}_2 &= \left\{ \widehat{\mathrm{KL}}(\{N_{i,j}(T)\}_{j \in \mathcal{O}'_i}) \leq \left(1 - \frac{\epsilon}{2}\right) \log T \right\}, \\
\mathcal{D}_{12} &= \mathcal{D}_1 \cap \mathcal{D}_2, \\
\mathcal{D}_{1 \setminus 2} &= \mathcal{D}_1 \cap \mathcal{D}_2^c.
\end{aligned}$$

First step ($\mathbb{P}\{\mathcal{D}_{12}\} = o(1)$): we have,

$$\begin{aligned}
&\mathbb{P}' \left(\mathcal{D}_{12} \cap \bigcap_{j_1, j_2 \in [K]} \{N_{j_1,j_2}(T) = n_{j_1,j_2}\} \right) \\
&= \int_{\mathcal{D}_{12} \cap \bigcap_{j_1,j_2 \in [K]} \{N_{j_1,j_2}(T) = n_{j_1,j_2}\}} \exp\left(-\widehat{\mathrm{KL}}(\{N_{i,j}(T)\}_{j \in \mathcal{O}'_i}) \right) d\mathbb{P} \\
&\geq \mathbb{E} \left[\mathbf{1} \left\{ \mathcal{D}_{12} \cap \bigcap_{j_1, j_2 \in [K]} \{N_{j_1,j_2}(T) = n_{j_1,j_2}\} \right\} \exp\left(-\left(1 - \frac{\epsilon}{2}\right) \log T \right) \right] \\
&= T^{-(1-\epsilon/2)} \mathbb{P} \left(\mathcal{D}_{12} \cap \bigcap_{j_1, j_2 \in [K]} \{N_{j_1,j_2}(T) = n_{j_1,j_2}\} \right).
\end{aligned}$$

summing over a disjoint union of events $\{\bigcap_{j_1,j_2 \in [K]} \{N_{j_1,j_2}(T) = n_{j_1,j_2}\}\}$ for each $j_1, j_2 \in \mathbb{N}$, we obtain

$$\mathbb{P}'(\mathcal{D}_{12}) \geq T^{-(1-\epsilon/2)} \mathbb{P}(\mathcal{D}_{12}).$$

By using this we have

$$\begin{aligned}\mathbb{P}(\mathcal{D}_{12}) &\le T^{(1-\epsilon/2)}\mathbb{P}'(\mathcal{D}_{12})\\ &\le T^{(1-\epsilon/2)}\mathbb{P}'\left\{N_{i,i}(T) < \sqrt{T}\right\}\\ &\le T^{(1-\epsilon/2)}\mathbb{P}'\left\{T - N_{i,i}(T) > T - \sqrt{T}\right\}\\ &\le T^{(1-\epsilon/2)}\frac{\mathbb{E}'[T - N_{i,i}(T)]}{T-\sqrt{T}} \qquad \text{(by the Markov inequality).} \end{aligned} \tag{6.8}$$

Since this algorithm is strongly consistent, $\mathbb{E}'[T - N_{i,i}(T)] \to o(T^a)$ for any $a > 0$. Therefore, the RHS of the last line of (6.8) is $o(T^{a-\epsilon/2})$, which, by choosing sufficiently small a, converges to zero as $T \to \infty$. In summary, $\mathbb{P}\{\mathcal{D}_{12}\} = o(1)$.

Second step ($\mathbb{P}\{\mathcal{D}_{1\setminus 2}\} = o(1)$): we have

$$\begin{aligned}&\mathbb{P}\{\mathcal{D}_{1\setminus 2}\}\\ &= \mathbb{P}\left\{\sum_{j\in\mathcal{O}_i'} N_{i,j}(T)d(\mu_{i,j},\mu_{i,j}') < (1-\epsilon)\log T, N_{i,i}(T) < \sqrt{T}, \sum_{j\in\mathcal{O}_i'}\widehat{\mathrm{KL}}_j(N_{i,j}(T)) > \left(1-\frac{\epsilon}{2}\right)\log T\right\}\\ &\le \mathbb{P}\left\{\max_{\{n_j\}\in\mathbb{N}^{|\mathcal{O}_i'|}, \sum_{j\in\mathcal{O}_i'} n_j d(\mu_{i,j},\mu_{i,j}')<(1-\epsilon)\log T} \sum_{j\in\mathcal{O}_i'}\widehat{\mathrm{KL}}_j(n_j) > \left(1-\frac{\epsilon}{2}\right)\log T\right\}.\end{aligned}$$

Note that

$$\max_{1\le n_j\le N}\widehat{\mathrm{KL}}_j(n_j) = \max_{1\le n_j\le N}\sum_{m=1}^{n_j}\log\left(\frac{\widehat{X}_{i,j}^m\mu_{i,j} + (1-\widehat{X}_{i,j}^m)(1-\mu_{i,j})}{\widehat{X}_{i,j}^m\mu_{i,j}' + (1-\widehat{X}_{i,j}^m)(1-\mu_{i,j}')}\right),$$

is the maximum of the sum of positive-mean random variables, and thus converges to is average (c.f., Lemma 10.5 in Bubeck, 2010). Namely,

$$\lim_{N\to\infty}\max_{1\le n_j\le N}\frac{\widehat{\mathrm{KL}}_j(n_j)}{N} = d(\mu_{i,j},\mu_{i,j}') \quad \text{a.s.} \tag{6.9}$$

Let $\delta > 0$ be sufficiently small. We have,

$$\begin{aligned}&\frac{\max_{\{n_j\}\in\mathbb{N}^{|\mathcal{O}_i'|}, \sum_{j\in\mathcal{O}_i'} n_j d(\mu_{i,j},\mu_{i,j}')<(1-\epsilon)\log T}\sum_{j\in\mathcal{O}_i'}\widehat{\mathrm{KL}}_j(n_j)}{\log T}\\ &\le \frac{\max_{\{n_j\}\in\mathbb{N}^{|\mathcal{O}_i'|}, \sum_{j\in\mathcal{O}_i':n_j>\delta\log T} n_j d(\mu_{i,j},\mu_{i,j}')<(1-\epsilon)\log T}\sum_{j\in\mathcal{O}_i'}\widehat{\mathrm{KL}}_j(n_j)}{\log T} + \frac{\delta K}{\min_{j\in\mathcal{O}_i'} d(\mu_{i,j},\mu_{i,j}')}.\end{aligned}$$

Combining this with the fact that (6.9) holds for any j, we have

$$\limsup_{N\to\infty}\frac{\max_{\{n_j\}\in\mathbb{N}^{|\mathcal{O}_i'|}, \sum_{j\in\mathcal{O}_i':n_j>\delta\log T} n_j d(\mu_{i,j},\mu_{i,j}')<(1-\epsilon)\log T}\sum_{j\in\mathcal{O}_i'}\widehat{\mathrm{KL}}_j(n_j)}{\log T} \le 1-\epsilon \quad \text{a.s.},$$

and thus

$$\limsup_{T\to\infty}\frac{\max_{\{n_j\}\in\mathbb{N}^{|\mathcal{O}_i'|}, \sum_{j\in\mathcal{O}_i'} n_j d(\mu_{i,j},\mu_{i,j}')<(1-\epsilon)\log T}\sum_{j\in\mathcal{O}_i'}\widehat{\mathrm{KL}}_j(n_j)}{\log T} \le 1-\epsilon+\widetilde{\mu}(\delta) \quad \text{a.s.} \tag{6.10}$$

By using the fact that (6.10) holds almost surely for any sufficiently small $\delta > 0$ and $1 - \epsilon/2 > 1 - \epsilon$, we have

$$\mathbb{P}\left(\max_{\{n_j\}\in\mathbb{N}^{|\mathcal{O}_i'|}, \sum_{j\in\mathcal{O}_i'} n_j d(\mu_{i,j},\mu_{i,j}')<(1-\epsilon)\log T} \sum_{j\in\mathcal{O}_i'} \widehat{\mathrm{KL}}_j(n_j) > \left(1-\frac{\epsilon}{2}\right)\log T\right) = o(1).$$

In summary, we obtain $\mathbb{P}\left\{\mathcal{D}_{1\setminus 2}\right\} = o(1)$.

Last step: we here have

$$\begin{aligned}
\mathcal{D}_1 &= \left\{\sum_{j\in\mathcal{O}_i'} N_{i,j}(T) d(\mu_{i,j},\mu_{i,j}') < (1-\epsilon)\log T\right\} \cap \left\{N_{i,i}(T) < \sqrt{T}\right\} \\
&= \left\{\sum_{j\in\mathcal{O}_i'} N_{i,j}(T)(d(\mu_{i,j},1/2)+\epsilon) < (1-\epsilon)\log T\right\} \cap \left\{N_{i,i}(T) < \sqrt{T}\right\} \qquad \text{(By (6.7))} \\
&\supseteq \left\{\sum_{j\in\mathcal{O}_i'} N_{i,j}(T)(d(\mu_{i,j},1/2)+\epsilon) + \frac{(1-\epsilon)\log T}{\sqrt{T}} N_{i,i}(T) < (1-\epsilon)\log T\right\},
\end{aligned}$$

where we used the fact that $\{A < C\} \cap \{B < C\} \supseteq \{A + B < C\}$ for $A, B > 0$ in the last line. Note that, by using the result of the previous steps, $\mathbb{P}\{\mathcal{D}_1\} = \mathbb{P}\{\mathcal{D}_{12}\} + \mathbb{P}\{\mathcal{D}_{1\setminus 2}\} = o(1)$. By using the complementary of this fact,

$$\mathbb{P}\left\{\sum_{j\in\mathcal{O}_i'} N_{i,j}(T)(d(\mu_{i,j},1/2)+\epsilon) + \frac{(1-\epsilon)\log T}{\sqrt{T}} N_{i,i}(T) \geq (1-\epsilon)\log T\right\} \geq \mathbb{P}\{\mathcal{D}_1^c\} = 1 - o(1).$$

Using the Markov inequality yields

$$\mathbb{E}\left\{\sum_{j\in\mathcal{O}_i'} N_{i,j}(T)(d(\mu_{i,j},1/2)+\epsilon) + \frac{(1-\epsilon)\log T}{\sqrt{T}} N_{i,i}(T)\right\} \geq (1-\epsilon)(1-o(1))\log T. \quad (6.11)$$

Because $\mathbb{E}[N_{i,i}(T)]$ is subpolynomial as a function of T due to the consistency, the second term in LHS of (6.11) is $o(1)$ and thus negligible. Lemma 27 follows from the fact that (6.11) holds for sufficiently small ϵ. □

6.8.2 Proof of Theorem 28

Proof. Proof of Theorem 28 We have

$$\begin{aligned}
\mathrm{Reg}(T) &= \frac{1}{2}\sum_{i\in[K]}\sum_{j\in[K]\setminus\{i\}}\frac{\Delta_{1,i}+\Delta_{1,j}}{2}N_{i,j}(T) + \sum_{i\in[K]}\frac{\Delta_{1,i}+\Delta_{1,i}}{2}N_{i,i}(T) \\
&\geq \sum_{i,j\in[K]:\mu_{i,j}<1/2}\frac{\Delta_{1,i}+\Delta_{1,j}}{2}N_{i,j}(T) + \sum_{i\in[K]}\frac{\Delta_{1,i}+\Delta_{1,i}}{2}N_{i,i}(T) \\
&\geq \sum_{i\in[K]\setminus\{1\}}\sum_{j\in\mathcal{O}_i}\frac{\Delta_{1,i}+\Delta_{1,j}}{2}N_{i,j}(T) \\
&= \sum_{i\in[K]\setminus\{1\}}\sum_{j\in\mathcal{O}_i}\frac{\Delta_{1,i}+\Delta_{1,j}}{2d(\mu_{i,j},1/2)}d(\mu_{i,j},1/2)N_{i,j}(T).
\end{aligned}$$

Taking the expectation on both sides and using Lemma 27 yield

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \sum_{i\in[K]\setminus\{1\}}\min_{j\in\mathcal{O}_i}\frac{\Delta_{1,i}+\Delta_{1,j}}{2d(\mu_{i,j},1/2)}(1-o(1))\log T.$$

□

6.9 Proof of Lemma 31

Proof. Proof of Lemma 31

This lemma essentially states that, the expected number of the rounds in which arm 1 is underestimated is $O(1)$. We show this by bounding the expected number of rounds before arm 1 is compared, for each fixed set of $\{N_{1,s}(t)\}$ and summing over $\{N_{1,s}(t)\}$. This technique is inspired by Lemma 16 in Honda and Takemura [2010]. Note that

$$\mathcal{U}^c(t) = \bigcup_{S\in 2^{[K]\setminus\{1\}}\setminus\{\emptyset\}}\left\{\bigcap_{s\in S}\{\widehat{\mu}_{1,s}(t)\leq 1/2\}\cap\bigcap_{s\notin S}\{\widehat{\mu}_{1,s}(t)>1/2\}\right\}. \tag{6.12}$$

Now we bound the number of rounds that the event

$$\bigcap_{s\in S}\{\widehat{\mu}_{1,s}(t)\leq 1/2\}\cap\bigcap_{s\notin S}\{\widehat{\mu}_{1,s}(t)>1/2\}$$

occurs. Let $\mathbb{N}$ be the set of non-zero natural numbers, $n_s\in\mathbb{N}$ and $x_s\in[0,\log 2]$ be arbitrary for each $s\in S$. Let $\widehat{\mu}^n_{i,j}$ be the empirical estimate of $\mu_{i,j}$ at n-th draw of pair (i,j). If $\{\widehat{\mu}^{n_s}_{1,s}\leq 1/2,\ d^+(\widehat{\mu}^{n_s}_{1,s},1/2)=x_s,\ N_{1,s}(t)=n_s\}$ holds for $s\in S$ and $\widehat{\mu}_{1,s}(t)>1/2$ holds for $s\notin S$ then

$$\mathcal{D}_1(t) = \sum_{s\in S}n_s d^+(\widehat{\mu}_{1,s}(t),1/2)$$

and therefore $\mathcal{J}_1(t)$ holds for any

$$t \geq \exp\left(\sum_{s\in S} n_s d^+(\widehat{\mu}_{1,s}(t), 1/2)\right). \tag{6.13}$$

If $\mathcal{J}_1(t)$ occurs, then arm 1 is in L_N of the next loop, and thus for some $s \in S$, $N_{1,s}$ is incremented within K rounds. Therefore we have

$$\sum_{t=T_{\text{init}}+1}^{T} \mathbf{1}\left[\bigcap_{s\in S}\{\widehat{\mu}_{1,s}(t) \leq 1/2,\, N_{1,s}(t) = n_s\} \cap \bigcap_{s\notin S}\{\widehat{\mu}_{1,s}(t) > 1/2\}\right] \leq \exp\left(\sum_{s\in S} n_s d^+(\widehat{\mu}_{1,s}^{n_s}, 1/2)\right) + K.$$

Letting $P_s(x_s) = \Pr[\widehat{\mu}_{1,s}^{n_s} \leq 1/2, d^+(\widehat{\mu}_{1,s}^{n_s}, 1/2) \geq x_s]$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=T_{\text{init}}+1}^{T} \mathbf{1}\left[\bigcap_{s\in S}\{\widehat{\mu}_{1,s}(t) \leq 1/2,\, N_{1,s}(t) = n_s\} \cap \bigcap_{s\notin S}\{\widehat{\mu}_{1,s}(t) > 1/2\}\right]\right] \\
&= \int_{\{x_s\}\in[0,\log 2]^{|S|}} \left(\exp\left(\sum_{s\in S} n_s x_s\right) + K\right) \prod_{s\in S} \mathrm{d}(-P_s(x_s)) \\
&= K\prod_{s\in S} P_s(0) + \prod_{s\in S}\int_{x_s\in[0,\log 2]} \mathrm{e}^{n_s x_s}\mathrm{d}(-P_s(x_s)) \\
&= K\prod_{s\in S} P_s(0) + \prod_{s\in S}\left([-\mathrm{e}^{n_s x_s}P_s(x_s)]_0^{\log 2} + \int_{x_s\in[0,\log 2]} n_s \mathrm{e}^{n_s x_s} P_s(x_s)\mathrm{d}x_s\right) \\
&\qquad \text{(integration by parts)} \\
&\leq (1+K)\prod_{s\in S} P_s(0) + \prod_{s\in S}\int_{x_s\in[0,\log 2]} n_s \mathrm{e}^{n_s x_s}\mathrm{e}^{-n_s(x_s + C_1(\mu_{1,s},1/2))}\mathrm{d}x_s \\
&\qquad \text{(by the Chernoff bound and Fact 36, where } C_1(\mu,\mu_2) = (\mu-\mu_2)^2/(2\mu(1-\mu_2))) \\
&\leq (1+K)\prod_{s\in S} \mathrm{e}^{-n_s d(1/2,\mu_{1,s})} + \prod_{s\in S}\int_{x_s\in[0,\log 2]} n_s \mathrm{e}^{-n_s C_1(\mu_{1,s},1/2)}\mathrm{d}x_s \\
&= (1+K)\prod_{s\in S} \mathrm{e}^{-n_s d(1/2,\mu_{1,s})} + \prod_{s\in S}(\log 2) n_s \mathrm{e}^{-n_s C_1(\mu_{1,s},1/2)}.
\end{aligned} \tag{6.14}$$

By summing (6.14) over $\{n_s\}$,

$$\begin{aligned}
&\sum_{t=T_{\text{init}}+1}^{T} \mathbb{P}\left[\bigcap_{s\in S}\{\widehat{\mu}_{1,s}(t) \leq 1/2\} \cap \bigcap_{s\notin S}\{\widehat{\mu}_{1,s}(t) > 1/2\}\right] \\
&\leq \sum_{\{n_s\}\in\mathbb{N}^{|S|}} \cdots \sum \left((1+K)\prod_{s\in S} \mathrm{e}^{-n_s d(1/2,\mu_{1,s})} + \prod_{s\in S}(\log 2) n_s \mathrm{e}^{-n_s C_1(\mu_{1,s},1/2)}\right) \\
&\leq (1+K)\prod_{s\in S}\frac{1}{\mathrm{e}^{d(1/2,\mu_{1,s})}-1} + (\log 2)^{|S|}\prod_{s\in S}\frac{\mathrm{e}^{C_1(\mu_{1,s},1/2)}}{(\mathrm{e}^{C_1(\mu_{1,s},1/2)}-1)^2},
\end{aligned}$$

where we used the fact that $\sum_{n=1}^{\infty} \mathrm{e}^{-nx} = 1/(\mathrm{e}^x + 1)$ and $\sum_{n=1}^{\infty} n\mathrm{e}^{-nx} = \mathrm{e}^x/(\mathrm{e}^x + 1)^2$. Using (6.12) and the union bound over all $S \in 2^{[K]\setminus\{1\}} \setminus \{\emptyset\}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}\{\mathcal{U}^c(t)\}\right] \\
&< (1+K) \prod_{s\in[K]\setminus\{1\}} \left(1 + \frac{1}{\mathrm{e}^{d(1/2,\mu_{1,s})} - 1}\right) + (\log 2)^{K-1} \prod_{s\in[K]\setminus\{1\}} \left(1 + \frac{\mathrm{e}^{C_1(\mu_{1,s},1/2)}}{(\mathrm{e}^{C_1(\mu_{1,s},1/2)} - 1)^2}\right) \\
&= O(\mathrm{e}^{AK}),
\end{aligned}$$

where $A = \log\left\{\max_{s\in[K]\setminus\{1\}} \max\left(1 + \frac{1}{\mathrm{e}^{d(1/2,\mu_{1,s})}-1}, \log 2\left(1 + \frac{\mathrm{e}^{C_1(\mu_{1,s},1/2)}}{(\mathrm{e}^{C_1(\mu_{1,s},1/2)}-1)^2}\right)\right)\right\}$. □

6.10 Proof of Lemma 32

Proof. Proof of Lemma 32

Except for the first loop, arm i must put into L_N before $\{l(t) = i\}$. For $t \geq T_{\mathrm{init}} + K + 1$ (i.e., after the first loop), let $\tau(t) < t$ be the round in the previous loop in which arm $l(t)$ is put into L_N. In the round, $\mathcal{J}_{l(t)}(\tau(t))$ is satisfied. With this definition, for any two rounds $t_1, t_2 \geq T_{\mathrm{init}} + K + 1$ such that $l(t_1) = l(t_2) = i$, $t_1 \neq t_2 \Rightarrow \tau(t_1) \neq \tau(t_2)$ holds because $\tau(t_1)$ and $\tau(t_2)$ belong to different loops. By using $\tau(t)$, we obtain

$$\begin{aligned}
&\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}[l(t) = i, N_{i,j}(t) \geq N_{i,j}^{\mathrm{Suf}}(\delta)] \\
&\leq K + \sum_{t=T_{\mathrm{init}}+K+1}^{T} \mathbf{1}[l(t) = i, \mathcal{U}^c(\tau(t))] + \sum_{t=T_{\mathrm{init}}+K+1}^{T} \mathbf{1}[l(t) = i, \mathcal{U}(\tau(t)), N_{i,j}(t) \geq N_{i,j}^{\mathrm{Suf}}(\delta)] \\
&\leq K + \sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}[\mathcal{U}^c(t)] + \sum_{t=T_{\mathrm{init}}+K+1}^{T} \mathbf{1}[l(t) = i, \mathcal{U}(\tau(t)), N_{i,j}(t) \geq N_{i,j}^{\mathrm{Suf}}(\delta)].
\end{aligned}$$

Note that the expectation of term $\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}[\mathcal{U}^c(t)]$ is bounded by Lemma 31. Between $\tau(t)$ and t, the only round in which pair (i, j) can be compared is the round of $\{l(t) = j\}$ that occurs at most once, and thus $N_{i,j}(t) - N_{i,j}(\tau(t)) \leq 1$. By using this fact, we obtain

$$\begin{aligned}
&\sum_{t=T_{\mathrm{init}}+K+1}^{T} \mathbf{1}[l(t) = i, \mathcal{U}(\tau(t)), N_{i,j}(t) \geq N_{i,j}^{\mathrm{Suf}}(\delta)] \\
&\leq \sum_{t=T_{\mathrm{init}}+K+1}^{T} \mathbf{1}[l(t) = i, \mathcal{J}_i(\tau(t)), \mathcal{U}(\tau(t)), N_{i,j}(\tau(t)) \geq N_{i,j}^{\mathrm{Suf}}(\delta) - 1] \\
&\leq \sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}[\mathcal{J}_i(t), \mathcal{U}(t), N_{i,j}(t) \geq N_{i,j}^{\mathrm{Suf}}(\delta) - 1].
\end{aligned}$$

We can bound this term via $\widehat{D}_i(t)$ as

$$\begin{aligned}
&\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}[\mathcal{J}_i(t), \mathcal{U}(t), N_{i,j}(t) \geq N_{i,j}^{\mathrm{Suf}}(\delta) - 1] \\
&\leq \sum_{n=\lceil N_{i,j}^{\mathrm{Suf}}(\delta)-1 \rceil}^{T} \mathbf{1}\left[\bigcup_{t=T_{\mathrm{init}}+1}^{T} \left(\widehat{D}_j(t) \leq \log t + f(K), N_{i,j}(t) = n \right) \right] \quad (\text{by } \mathcal{U}(t) \Rightarrow \widehat{D}_1(t) = 0) \\
&\leq \sum_{n=\lceil N_{i,j}^{\mathrm{Suf}}(\delta)-1 \rceil}^{T} \mathbf{1}\left[\bigcup_{t=T_{\mathrm{init}}+1}^{T} \left(N_{i,j}(t) = n, N_{i,j}(t) d^+(\widehat{\mu}_{i,j}^n, 1/2) \leq \log t + f(K) \right) \right] \\
&\leq \sum_{n=\lceil N_{i,j}^{\mathrm{Suf}}(\delta)-1 \rceil}^{T} \mathbf{1}\left[(N_{i,j}^{\mathrm{Suf}}(\delta) - 1) d^+(\widehat{\mu}_{i,j}^n, 1/2) \leq \log T + f(K) \right] \\
&\leq \sum_{n=\lceil N_{i,j}^{\mathrm{Suf}}(\delta)-1 \rceil}^{T} \mathbf{1}\left[d^+(\widehat{\mu}_{i,j}^n, 1/2) \leq \frac{d(\mu_{i,j}, 1/2)}{1+\delta} \right].
\end{aligned}$$

Therefore, by letting $\mu \in (1/2, \mu_{i,j})$ be a real number such that $d(\mu, 1/2) = \frac{d(\mu_{i,j}, 1/2)}{1+\delta}$, we obtain from the Chernoff bound and the monotonicity of $d^+(\cdot, 1/2)$ that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}[\mathcal{J}_i(t), \mathcal{U}(t), N_{i,j}(t) \geq N_{i,j}^{\mathrm{Suf}}(\delta) - 1] \right] &\leq \sum_{n=\lceil N_{i,j}^{\mathrm{Suf}}(\delta)-1 \rceil}^{T} \mathbb{P}\left[d^+(\widehat{\mu}_{i,j}^n, 1/2) \leq \frac{d(\mu_{i,j}, 1/2)}{1+\delta} \right] \\
&\leq \sum_{n=\lceil N_{i,j}^{\mathrm{Suf}}(\delta)-1 \rceil}^{T} \exp\left(-d(\mu, \mu_{i,j}) n \right) \\
&\leq \frac{1}{\exp\left(d(\mu, \mu_{i,j}) \right) - 1} < \frac{1}{d(\mu, \mu_{i,j})}.
\end{aligned}$$

From the Pinsker's inequality it is easy to confirm that $d(\mu, \mu_{i,j}) = \Omega(\delta^2)$, which completes the proof. □

6.11 Optimal Regret Bound: Full Proof of Theorem 30

Proof. Proof of Theorem 30

Events: define

$$\mathcal{A}_i = \bigcap_{i,j \in [K]} \{ |\widehat{\mu}_{i,j}^{\lceil \alpha \log \log T \rceil} - \mu_{i,j}| < \Delta_i^{\mathrm{suf}} \}$$

for sufficiently small but fixed $\Delta_i^{\mathrm{suf}} > 0$. It is easy to see from the condinuity of $d^+(\mu_{i,j}, 1/2)$ in $\mu_{i,j}$ that $\mathcal{A}_i$ implies $\widehat{b}^\star(i) = b^\star(i)$ when we let $\Delta_i^{\mathrm{suf}} > 0$ be sufficiently small with respect to $\{\mu_{i,j}\}_{i,j \in [K]}$. Let also

$$\mathcal{B}_i(t) = \{ \widehat{\mu}_{i, b^\star(i)}(t) < 1/2 \}.$$

First step (regret decomposition): like RMED1, in RMED2FH $\mathbb{E}[\mathcal{U}(t)]$ holds with high probability (i.e., Lemma 31). In the following, we bound the regret under $\mathcal{U}(t)$: let

$$\begin{aligned} r_i(t) &= \mathbf{1}\{l(t)=i, \mathcal{U}(t)\} r(t) \\ &= \underbrace{\mathbf{1}\{l(t)=i, \mathcal{U}(t), \mathcal{A}_i, \mathcal{B}_i(t)\} r(t)}_{\text{(A)}} + \underbrace{\mathbf{1}\{l(t)=i, \mathcal{U}(t), \{\mathcal{A}_i^c \cup \mathcal{B}_i^c(t)\}\} r(t).}_{\text{(B)}} \end{aligned} \tag{6.15}$$

In the following, we first bound the terms (A) and (B), and then summarizing all terms to prove Theorem 30.

Second step (bounding (A)): note that, $\{l(t)=i, \mathcal{U}(t), \mathcal{A}_i, \mathcal{B}_i(t)\}$ is a sufficient condition for $\widehat{b}^\star(i) = b^\star(i)$ and $\widehat{b}^\star(i) \in \widehat{\mathcal{O}}_i(t)$. Therefore,

$$\begin{aligned} &\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}\{l(t)=i, \mathcal{U}(t), \mathcal{A}_i, \mathcal{B}_i(t)\} r(t) \\ &\leq \sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}\{l(t)=i, N_{i,b^\star(i)}(t) \geq N^{\mathrm{Suf}}_{i,b^\star(i)}(\delta)\} + \frac{\Delta_{1,i}+\Delta_{1,b^\star(i)}}{2} N^{\mathrm{Suf}}_{i,b^\star(i)}(\delta) + \frac{\Delta_{1,i}}{2} \frac{N^{\mathrm{Suf}}_{i,b^\star(i)}(\delta)}{\log\log T}. \end{aligned}$$

By applying Lemma 32 with $j = b^\star(i)$, for sufficiently small $\delta > 0$ we have

$$\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}\{l(t)=i, N_{i,b^\star(i)}(t) \geq N^{\mathrm{Suf}}_{i,b^\star(i)}(\delta)\}\right] \leq O\left(\frac{1}{\delta^2}\right).$$

In summary, term (A) is bounded as

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T} \mathbf{1}\{l(t)=i, \mathcal{U}(t), \mathcal{A}_i, \mathcal{B}_i(t)\} r(t)\right] \\ &\qquad \leq \frac{\Delta_{1,i}+\Delta_{1,b^\star(i)}}{2} N^{\mathrm{Suf}}_{i,b^\star(i)}(\delta) + O\left(\frac{\log T}{\log\log T}\right) + O\left(\frac{1}{\delta^2}\right). \end{aligned} \tag{6.16}$$

Third step (bounding (B)): now we consider the case $\{l(t)=i, \mathcal{U}(t), \{\mathcal{A}_i^c \cup \mathcal{B}_i^c(t)\}\}$. Under this event $\widehat{b}^\star(i) = b^\star(i)$ does not always hold but we can see that $m(t) \in \{\widehat{b}^\star(i), 1\}$ still holds. Furthermore, under this event arm $\widehat{b}^\star(i)$ is selected as $m(t)$ at most $(\log\log T) N_{i,1}(T) + 1$ times due to Line 12 of Algorithm 12. By using these facts, we

have,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T}\mathbf{1}\{l(t)=i,\mathcal{U}(t),\{\mathcal{A}_i^c\cup\mathcal{B}_i^c(t)\}\}r(t)\right]\\
&\le\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T}\mathbf{1}\{l(t)=i,\mathcal{U}(t),\{\mathcal{A}_i^c\cup\bigcup_{t'=T_{\mathrm{init}}+1}^{T}\mathcal{B}_i^c(t')\}\}\right]\\
&\le\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}^{T}\mathbf{1}\{l(t)=i,N_{i,1}(t)\ge N_{i,1}^{\mathrm{Suf}}(\delta)\}\right]\\
&\qquad+\mathbb{P}\left\{\mathcal{A}_i^c\cup\bigcup_{t'=T_{\mathrm{init}}+1}^{T}\mathcal{B}_i^c(t')\right\}\left(N_{i,1}^{\mathrm{Suf}}(\delta)\log\log T+1+N_{i,1}^{\mathrm{Suf}}(\delta)\right)\\
&\le O\left(\frac{1}{\delta^2}\right)+\mathbb{P}\left\{\mathcal{A}_i^c\cup\bigcup_{t'=T_{\mathrm{init}}+1}^{T}\mathcal{B}_i^c(t')\right\}O\left(N_{i,1}^{\mathrm{Suf}}(\delta)\log\log T\right)\\
&\qquad\text{(by Lemma 32).}
\end{aligned}$$

The following lemma bounds $\mathbb{P}\left\{\mathcal{A}_i^c\cup\bigcup_{t'=T_{\mathrm{init}}+1}^{T}\mathcal{B}_i^c(t')\right\}$.

Lemma 33. *For RMED2FH, there exists $C_2=C_2(\{\mu_{i,j}\},K,\alpha)>0$ such that*

$$\mathbb{P}\left\{\mathcal{A}_i^c\cup\bigcup_{t=T_{\mathrm{init}}+1}^{T}\mathcal{B}_i^c(t)\right\}=O((\log T)^{-C_2}).$$

In summary, term (B) is bounded as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=T_{\mathrm{init}}+1}\mathbf{1}\{l(t)=i,\mathcal{U}(t),\{\mathcal{A}_i^c\cup\mathcal{B}_i^c(t)\}\}r(t)\right]\\
&\qquad\le O\left(\frac{1}{\delta^2}\right)+O\left(N_{i,1}^{\mathrm{Suf}}(\delta)(\log T)^{-C_2}\log\log T\right).\quad(6.17)
\end{aligned}$$

Last step (regret bound):

$$\mathbb{E}[\mathrm{Reg}(T)] \le T_{\mathrm{init}} + \sum_{t=T_{\mathrm{init}}+1}^{T} \left(\mathbb{P}\{\mathcal{U}^c(t)\} + \sum_{i\in[K]\setminus\{1\}} \mathbb{P}\{\mathcal{U}(t), l(t)=i\} r_i(t) \right)$$

$$\le T_{\mathrm{init}} + \sum_{t=T_{\mathrm{init}}+1}^{T} \left(O(1) + \sum_{i\in[K]\setminus\{1\}} \mathbb{P}((\mathrm{A})+(\mathrm{B})) \right) \quad \text{(by Lemma 31 and inequality (6.15))}$$

$$\le O(\alpha K^2 \log\log T)$$

$$+ \sum_{i\in[K]\setminus\{1\}} \left\{ \frac{\Delta_{1,i}+\Delta_{1,b^\star(i)}}{2} N^{\mathrm{Suf}}_{i,b^\star(i)}(\delta) + O\left(\frac{\log T}{\log\log T}\right) \right.$$

$$\left. + O\left(\frac{1}{\delta^2}\right) + O\left(N^{\mathrm{Suf}}_{i,1}(\delta)(\log T)^{-C_2}\log\log T\right) \right\}$$

(by (6.16) and (6.17))

$$\le O(\alpha K^2 \log\log T) + O(1) + \sum_{i\in[K]\setminus\{1\}} \frac{(\Delta_{1,i}+\Delta_{1,b^\star(i)})((1+\delta)\log T)}{2d(\mu_{i,b^\star(i)}, 1/2)}$$

$$+ O\left(\frac{K\log T}{\log\log T}\right) + O\left(\frac{K}{\delta^2}\right) + O\left(K(\log T)^{1-C_2}\log\log T\right) + O\left(Kf(K)\right). \tag{6.18}$$

Combining (6.18) with the fact that $O\left(K(\log T)^{1-C_2}\log\log T\right) = o\left(\frac{K\log T}{\log\log T}\right)$ completes the proof. □

6.11.1 Proof of Lemma 33

Proof. Proof of Lemma 33
We bound $\mathbb{P}\{\mathcal{A}_i^c\}$ and $\mathbb{P}\{\bigcup_{t=T_{\mathrm{init}}+1}^{T} \mathcal{B}_i^c(t)\}$ separately. On the one hand,

$$\mathbb{P}\{\mathcal{A}_i^c\} = \mathbb{P}\left\{ \bigcup_{i,j\in[K]} |\widehat{\mu}_{i,j}^{\lceil \alpha\log\log T\rceil} - \mu_{i,j}| \ge \Delta_i^{\mathrm{suf}} \right\} \le \sum_{i,j\in[K]} \mathbb{P}\{|\widehat{\mu}_{i,j}^{\lceil \alpha\log\log T\rceil} - \mu_{i,j}| \ge \Delta_i^{\mathrm{suf}}\}$$

$$\le \sum_{i,j\in[K]} 2\exp\left(-2(\Delta_i^{\mathrm{suf}})^2 \alpha \log\log T\right) \quad \text{(by the Chernoff bound and Pinsker's inequality)}$$

$$= \sum_{i,j\in[K]} 2\left(\log T\right)^{-2(\Delta_i^{\mathrm{suf}})^2\alpha} = 2K^2\left(\log T\right)^{-2(\Delta_i^{\mathrm{suf}})^2\alpha} = O((\log T)^{-C_a}),$$

where $C_a = 2(\Delta_i^{\mathrm{suf}})^2\alpha/K^2 > 0$. On the other hand,

$$
\begin{aligned}
&\mathbb{P}\left\{\bigcup_{t=T_{\mathrm{init}}+1}^{T}\mathcal{B}_i^c(t)\right\}\\
&=\mathbb{P}\left\{\bigcup_{t=T_{\mathrm{init}}+1}^{T}\widehat{\mu}_{i,b^\star(i)}(t)<1/2\right\}\le\mathbb{P}\left(\bigcup_{n=\lceil\alpha\log\log T\rceil}^{\infty}\{N_{i,b^\star(i)}(t)=n,\widehat{\mu}^n_{i,b^\star(i)}<1/2\}\right)\\
&\le\sum_{n=\lceil\alpha\log\log T\rceil}^{\infty}\mathbb{P}\{N_{i,b^\star(i)}(t)=n,\widehat{\mu}^n_{i,b^\star(i)}<1/2\}\\
&\le\sum_{n=\lceil\alpha\log\log T\rceil}^{\infty}\exp\left(-d(1/2,\mu_{i,b^\star(i)})n\right)\qquad\text{(by the Chernoff bound)}\\
&\le(\log T)^{-\alpha d(1/2,\mu_{i,b^\star(i)})}\sum_{n=0}^{\infty}\exp\left(-d(1/2,\mu_{i,b^\star(i)})n\right)\\
&\le(\log T)^{-\alpha d(1/2,\mu_{i,b^\star(i)})}\left(1+\frac{1}{d(1/2,\mu_{i,b^\star(i)})-1}\right)=O((\log T)^{-C_b}),
\end{aligned}
$$

where $C_b = \alpha d(1/2,\mu_{i,b^\star(i)}) > 0$. The proof is completed by letting $C_2 = \min(C_a, C_b)$ and taking the union bound of $\mathbb{P}\{\mathcal{A}_i^c\}$ and $\mathbb{P}\{\bigcup_{t=T_{\mathrm{init}}+1}^{T}\mathcal{B}_i^c(t)\}$. □

Chapter 7

Conclusions and Future Work

In this section, we present our conclusions on the work that makes up this thesis and discuss various extensions to the multi-armed bandit problem.

7.1 Concluding Remarks

In this thesis, we have discussed the multi-armed bandit problem and its extensions. In particular, the framework of the problem was described in Chapter 2 and 3. The study of the multi-armed bandit problem began in the statistics community, and the framework of the stochastic bandit was established through the introduction of an asymptotic regret lower bound for strongly consistent algorithms and an asymptotically optimal algorithm. The regret, which gives a criterion for balancing exploration and exploitation, is written in terms of the KL divergence between the true model and the other models in which the optimal arm is different from the one of the true model. There are many algorithms that are effective at solving the bandit problem. Among them, UCB, TS, and DMED are known to have asymptotically optimal regret bounds.

The stochastic bandit problem is a simple yet extensible framework. Motivated by its applications to web systems, we have rethought its three core notions: (i) sequential selection of arms, (ii) the criterion of selection, and (iii) reward feedback: namely, we have studied the lock-up restriction (Chapter 4), the multiple-play extension (Chapter 5), and the dueling extension (Chapter 6). We have shown that it is possible to balance exploration and exploitation in these extensions.

7.2 Other Directions

In this chapter, we discuss some of the other extensions of the bandit problem. The wide variety of extensions shows the flexibility of the bandit framework.

7.2.1 Continuous bandit problems

One version of the bandit problem, which we call a continuous bandit problem, has an infinite number of arms. Since the number of available samples is finite, we do not have enough time to check an infinite number of arms, and thus, some structural assumptions have to be placed on the arms. There is no unified solution to the continuous-armed bandit problem since the structure of the arms varies among problems.

Most studies assume the arms have a metric structure. Some papers, such as Agrawal [1995a], Kleinberg et al. [2008], Bubeck et al. [2011], and Magureanu et al. [2014], studied the case where the expected reward is Lipschitz as a function of the arm space. The case where the expected reward function is linear [Dani et al., 2008] or convex [Flaxman et al., 2005] has also been studied. Recently, Bayesian optimization [Mockus, 1974, Snoek et al., 2012], a global optimization over a continuous domain, has attracted increasing attention in the machine learning community for its application to tuning the hyperparameters of machine learning algorithms. In Bayesian optimization, the relation between arms and the reward function is often modeled as a Gaussian process. At each round, the forecaster selects a point in the Gaussian process and receives a (possibly noisy) observation of the point. Bayesian optimization can be considered to be a continuous bandit problem where the correlation between arms is represented by a Gaussian process.

Other than the metric structure, Yu and Mannor [2011] proposed a bandit problem whose parameters are restricted to be unimodal with respect to the associated graph structure. Furthermore, an optimal algorithm under the unimodal assumption has been proposed [Combes and Proutiere, 2014]. Some studies (e.g., Bubeck et al. [2011]) have generalized the metric-induced continuous bandit problems to a certain topological structure class.

In general, UCB, TS, and DMED can be extended to continuous bandit problems. Even though the number of arms is infinite, the arms are mutually dependent, and thus exploring the region of high uncertainty helps. Selecting an arm yields not only information on the arm but also information on one of the arms that are close to the selected one.

7.2.2 Many-armed bandit problems

An (infinitely) many-armed bandit problem is a version of the multi-armed bandit problem in which the number of arms is large or infinite. Unlike the continuous bandits, this setting does not assume there is any metric structure among the arms. If the number of samples is finite, identifying the optimal arm is not possible given the large number of arms. Interestingly, no-regret learning is possible given access to an infinite pool of arms and given some assumption on the tail probability of the distribution of arms. The stochastic many-armed bandit problem was first studied in a general setting by Mallows

and Robbins [1964]. In particular, the case of the Bernoulli bandit was studied by Herschkorn et al. [1996]. They proposed an asymptotically no-regret algorithm in which $\mathbb{E}[\mathrm{Reg}(T)]/T \to 0$.

Later, Berry et al. [1997] proposed an algorithm with $O(\sqrt{T})$ regret that is optimal up to a leading constant factor. The constant factor in the case of a known time horizon T was tightened by Bonald and Proutière [2013] to $\sqrt{2T}$. These studies assume the parameter μ_i of a new arm is uniformly distributed over $[0, 1]$. In general, the regret depends on the tail probability of the distribution of $\{\mu_i\}$ around its maximum. More general reward settings were studied by Wang et al. [2008]. In such settings, David and Shimkin [2014] studied the non-retainable case in which an algorithm must abandon an arm unless it is immediately in use.

UCB, TS, and DMED are not directly applicable to the many-armed bandit problems. Note that algorithms for the multi-armed bandit problems are designed to explore each arm $O(\log T)$ times. However, in a many-armed bandit problem, the number of rounds is smaller than the number of arms, and thus it is not possible to search all arms extensively. An algorithm for solving many-armed bandit problems needs to discard arms faster than the confidence bound shrinks.

Here, we explain the idea of the two-target algorithm devised by Bonald and Proutière [2013]. Let us consider a many-armed bandit problem with Bernoulli rewards. The parameter of each arm μ_i, $i = 1, 2, \dots$ is also a random variable that is uniformly distributed over $[0, 1]$. At each round $t = 1, 2, \dots$, the algorithm selects some arm $I(t) \in \mathbb{N}$ and receives the corresponding reward $\widehat{X}_t \sim \mathrm{Bernoulli}(\mu_{I(t)})$. The regret is defined as

$$\mathrm{Reg}(T) = T - \sum_{t=1}^{T} \widehat{X}_{I(t)}$$

because the optimal arm has an expectation of 1. The goal of the algorithm is to minimize the expectation of the regret. The two-target algorithm with a known horizon T continues exploring an arm until two targets l_1 and l_2 are reached. If they are reached, it exploits the current arm until the end of the rounds. The first target quickly discards a bad arm, and the second target carefully checks whether the arm is truly good or not. The algorithm involves a parameter $m \geq 2$, and the regret of the two-target algorithm with $l_1 = \lfloor (T/2)^{1/3} \rfloor$ and $l_2 = \lfloor m(T/2)^{1/2} \rfloor$ is asymptotically bounded as

$$\limsup_{T \to +\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\sqrt{T}} \leq \sqrt{2} + \frac{1}{m\sqrt{2}},$$

which matches the lower bound as $m \to \infty$.

7.2.3 Monte Carlo Tree Search

Consider an abstract game, such as Chess or Go. At each round, the player whose turn it is needs to decide the next move. In a game playing situation, the search space

of the sequence of player moves can be represented as a tree. In a two-player game, the child node of the root node (current state) is the current player' s move, and the child of that node is the opponent's move, and so on. The objective of a player is to find the exact value of each next move in the current situation, which can be done by searching the tree. Since the number of nodes of the tree grows exponentially with the depth of the tree, searching over the entire space is computationally prohibitive, and we need to use an ingenious algorithm to deal with this situation. The Monte Carlo tree search (MCTS) is a search algorithm on a tree structure involving randomization. When searching the tree, Upper Confidence Bound for Trees (UCT) [Kocsis and Szepesvári, 2006] selects the next move on a node based on its UCB index. This idea works well in tree searches, and as a result, many contemporary implementations of MCTS have been based on some variant of UCT. MCTS is especially successful in computer Go. An extensive survey on this topic is presented in Browne et al. [2012].

7.2.4 Use of contextual information

The contextual bandit problem [Langford and Zhang, 2007] is an extension of the standard multi-armed bandit that involves additional information. In this framework, the algorithm is informed of the side information before it selects an arm. In particular, in a content recommendation setting, the side information can be considered to be personal information such as demographic data, which correlates with the preference about the content.

Li et al. [2010] reported that the use of personal information as a context for news article recommendation on the Yahoo! homepage can increase the click-through rate of users by up to 12.5%. Agarwal et al. [2009] proposed a framework for bandit-based web content optimization that addressed several practical issues, such as delays and non-stationarities. Scott [2015] discussed the application of TS to web systems and proposed several extensions including a contextual version. These papers do not include regret analyses.

Technically, the stochastic analysis of contextual bandits is rather involved, because the algorithm selects an arm after it receives the side information. A martingale analysis based on the self-normalized bound [de la Peña et al., 2004] is often used in stochastic analyses of contextual bandit problems [Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011]. Alternatively, one can formalize the contextual bandit problems as a variant of the adversarial bandit; in fact, the paper that coined the word "contextual bandit" problem formalized it as a version of the adversarial bandit problem [Langford and Zhang, 2007].

7.2.5 Game theoretic framework

A major application of the multi-armed bandit problem is search engine advertising. When a user makes a query, relevant advertisements (ads) are listed in the search engine, which is called a broad-match procedure. The search engine chooses an ad among them and displays it on the search results page. If the ad is clicked by the user, the corresponding advertiser pays according to his/her bid, which is determined by the (generalized) second price mechanism. For more details on search engine advertising, see a recent review by Qin et al. [2014]. Given a user's query, maximization of the search engine' s revenue boils down to the standard multi-armed bandit problem. However, taking the advertisers' strategic bidding schedule into consideration, revenue maximization requires a game-theoretic analysis. This is because the search engine needs to motivate advertisers to bid truthfully in order to have a sound auction.

Truthful multi-armed bandit mechanisms [Babaioff et al., 2009] are a game-theoretic extension of the multi-armed bandit problem that models social welfare maximization on pay-per-click auctions with an unknown click-through rate. Babaioff et al. [2009] showed that in regard to the strategic activity of the advertisers, the optimal balance between the exploration and exploitation is different from the one of the standard stochastic bandit problem. Devanur and Kakade [2009] considered a similar problem from the search engine's view: they considered the revenue maximization problem and showed that there is some price that guarantees a strategy-proof mechanism that is robust to the strategic bidding of the advertisers.

Xu et al. [2013] studied a two-stage framework that models revenue optimization in search engine advertising. At the beginning, K advertisers submit bids $b_1, \ldots, b_K$. The first T_1 rounds are for learning the click-through rates (i.e., exploration). At each round t during this stage, an ad is selected by using a bandit algorithm. At the end of this stage, the empirical click-through rate $\widehat{\mu}_i$ of each arm is fixed. The remaining T_2 round, called the second price auction stage, is for exploitation. In this stage, the arm is fixed such that $\arg\max_i b_i \widehat{\mu}_i$, and the price per click is determined in accordance with the estimated second price value. Xu et al. [2013] showed that an algorithm with a logarithmic exploration rate, such as UCB1, will generate some bias that reduces the search engine's revenue. Namely, the expected revenue is smaller than the true second price. They proposed a simple bias-removing technique as a remedy for this problem: instead of keeping one history of each advertisement, one can keep two click-og histories by allocating two impressions at each round. This procedure increases the search engine's revenue even though the number of the rounds in the first stage is halved.

Moreover, Hummel and McAfee [2014] studied a pay-per-click auction in which the click-through rates of ads are unknown. They studied maximization of social welfare based on

the assumption that the value of the ads is drawn from some probability distribution and showed that the optimal exploration is $O(1/N_i(t)^2)$, which is smaller than the amount of exploration that is expected in the stochastic bandit formalization. Note that their model involves a discount factor $\delta < 1$, and thus strong consistency is not required.

7.2.6 Partial monitoring

Partial monitoring [Piccolboni and Schindelhauer, 2001, Bartók et al., 2011] is a wide class of problems that encompasses the multi-armed bandit problem. This problem involves actions and outcomes. At the beginning of each round, a learner selects an action and receives a signal that gives partial information on a stochastic outcome. The reward is a deterministic function of the selected action and the outcome, which, unlike the bandit problem, is not disclosed to the learner. The goal of the learner is to maximize the cumulative rewards over rounds. This problem is harder than the multi-armed bandit problem in the sense that the reward cannot be uniquely determined by the feedback. One of the seminal results on this problem is the classification of the distribution-independent regret, the worst-case regret over the model parameters. Bartók et al. [2011] classified the partial monitoring problems into four categories in terms of the distribution-independent regret: a trivial problem with zero regret, an easy problem with $\tilde{\Theta}(\sqrt{T})$ regret[*1], a hard problem with $\Theta(T^{2/3})$ regret, and a hopeless problem with $\Theta(T)$ regret. This shows that the class of partial monitoring problems is not limited to the bandit sort, but also includes larger classes of interesting problems, such as dynamic pricing. Contrary to these developments, not much is known on the distribution-dependent regret, which is the standard metric in the bandit problem. The (distribution-dependent) regret lower bound in the stochastic partial monitoring problem under strong consistency, which is an extension of the one in the multi-armed bandit problem, was recently derived by Komiyama et al. [2015c]. They also proposed PM-DMED, an asymptotically optimal algorithm for all learnable classes of stochastic partial monitoring problems. The result is strong in the sense that it entails an asymptotically optimal algorithm for solving the multi-armed bandit problem with binary rewards. One interesting direction for future work is to make the partial monitoring problem scalable: it is usually the case that the number of outcomes is exponentially large with respect to the number of actions. The existing algorithms for partial monitoring do not scale well in this case.

[*1] Note that $\tilde{\Theta}$ ignores a polylog factor.

Appendix A

Appendix

Fact 34. (Beta-Binomial equality) *Let $F^{\text{beta}}_{\alpha,\beta}(y)$ be the cdf of the beta distribution with integer parameters α and β. Let $F^{\text{B}}_{n,p}(\cdot)$ be the cdf of the binomial distribution with parameters n, p. Then,*

$$F^{\text{beta}}_{\alpha,\beta}(y) = 1 - F^{\text{B}}_{\alpha+\beta-1,y}(\alpha - 1),$$

Fact 35. (The Pinsker's inequality)
For $p, q \in (0,1)$, the KL divergence between two Bernoulli distributions is bounded as

$$d(p,q) \geq 2(p-q)^2.$$

Fact 36. (A minimum difference between divergences [Lemma 13 in Honda and Takemura, 2010])
For any μ and μ_2 satisfying $0 < \mu_2 < \mu < 1$. Let $C_1(\mu, \mu_2) = (\mu - \mu_2)^2/(2\mu(1-\mu_2))$. Let $d(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$. Then, for any $\mu_3 \leq \mu_2$,

$$d(\mu_3, \mu) - d(\mu_3, \mu_2) \geq C_1(\mu, \mu_2) > 0.$$

Acknowledgment

In this final part of the thesis, the author acknowledges the people without whom this thesis is completed.

First of all, the author greatly thanks the supervisor Hiroshi Nakagawa for his entire support. He generously accepted me to enter the Ph.D. course at the University of Tokyo. Before that, I was a software engineer and not very familiar with machine learning: he helped me to start my research on the machine learning by providing various topics. Moreover, he provided me a freedom of research throughout the Ph.D. course.

I acknowledge the members of the Ph.D. committee for their time and variable feedback.

The author thank Issei Sato for many discussions, inspiring ideas, and other efforts to help me. It was a great pleasure to be able to collaborate with him. My senior Hidekazu Oiwa influenced me a lot. His philosophy on information science research took an important role in developing the basis of my research. It was a great pleasure to be able to collaborate with him. Many thanks to Junya Honda, who is a collaborator of three papers about the multi-armed bandit problem. I owe him various ideas related to the efficiency of bandit algorithms. My mathematical skill was improved throughout working with him. I am grateful to Hisashi Kashima with whom I collaborated in the study of the dueling bandit problem.

During the Ph.D. period, I had three chances of collaboration with people outside the university. I thank Katsuhiko Ishiguro for mentoring my internship at NTT Communication Science Laboratory in August 2012. He taught me basics on nonparametric Bayesian and latent models. I would like to thank Tao Qin for mentoring my internship at Microsoft Research Asia from November 2013 to February 2014. The greatest thing I learned during the internship was the relationship between machine learning and game-theoretic frameworks. I thank Yasuyuki Sogawa for mentoring my internship at NEC Knowledge Discovery Research Laboratory from October 2014 to March 2015. I learned a lot about the reinforcement learning and Gaussian processes there.

I thank the Japan Society for the Promotion of Science (JSPS) for a grant that is necessary to complete this research.

I thank all Nakagawa laboratory members for sharing time with me. Sometimes, research is not very successful: the goal is hard to see, and the road that leads it can be long and winding. When you are in struggle, it is a great help to have friends with the

same goal in a long journey. Finally, I thank my family for their encouraging attitude and entire support.

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems 24: Proceedings of the 25th Annual Conference on Neural Information Processing Systems 2011 (NIPS 2011)*, pages 2312–2320, 2011.

Naoki Abe and Atsuyoshi Nakamura. Learning to Optimally Schedule Internet Banner Advertisements. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, pages 12–21, 1999.

Deepak Agarwal, Bee chung Chen, and Pradheep Elango. Explore/Exploit Schemes for Web Content Optimization. In *Proceedings of the Ninth IEEE International Conference on Data Mining (ICDM 2009)*, pages 1–10, 2009.

Gagan Aggarwal, Jon Feldman, S. Muthukrishnan, and Martin Pál. Sponsored Search Auctions with Markovian Users. In *Proceedings of the Internet and Network Economics, 4th International Workshop (WINE 2008)*, pages 621–628, 2008.

Rajeev Agrawal. The Continuum-Armed Bandit Problem. *SIAM J. Control Optim.*, 33 (6):1926–1951, November 1995a. ISSN 0363-0129.

Rajeev Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078, 1995b.

Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT 2012)*, pages 39.1–39.26, 2012.

Shipra Agrawal and Navin Goyal. Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, pages 99–107, 2013a.

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 127–135, 2013b.

Nir Ailon, Zohar Shay Karnin, and Thorsten Joachims. Reducing Dueling Bandits to Cardinal Bandits. In *Proceedings of the 31th International Conference on Machine Learning (ICML 2014)*, pages 856–864, 2014.

Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient

allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards. *Automatic Control, IEEE Transactions on*, 32(11):968–976, 1987.

Jean-Yves Audibert and Sébastien Bubeck. Minimax Policies for Adversarial and Stochastic Bandits. In *Proceedings of the 22nd Conference on Learning Theory (COLT 2009)*, 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19): 1876–1902, 2009.

Peter Auer, Nicoló Cesa-bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47:235–256, 2002a.

Peter Auer, Nicoló Cesa-bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a Rigged Casino: The Adversarial Multi-Arm Bandit Problem. In *IEEE Symposium on Foundations of Computer Science*, pages 322–331, 1995.

Peter Auer, Yoav Freund, and Robert E. Schapire. The non-stochastic multi-armed bandit problem. *Siam Journal on Computing*, 2002b.

Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 45–53, 2004.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms: extended abstract. In *ACM Conference on Electronic Commerce (EC 2009)*, pages 79–88, 2009.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax Regret of Finite Partial-Monitoring Games in Stochastic Environments. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT 2011)*, pages 133–154, 2011.

Richard Bellman. A Problem in the Sequential Design of Experiments. *Sankhya: The Indian Journal of Statistics*, 16(3/4):pp. 221–229, 1956. ISSN 00364452.

Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, and Larry A. Shepp. Bandit problems with infinitely many arms. *Ann. Statist.*, 25(5):2103–2116, 10 1997.

Donald A. Berry and Bert Fristedt. *Bandit problems : sequential allocation of experiments / Donald A. Berry, Bert Fristedt.* Chapman and Hall London ; New York, 1985. ISBN 0412248107.

Thomas Bonald and Alexandre Proutière. Two-Target Algorithms for Infinite-Armed Bandits with Bernoulli Rewards. In *Advances in Neural Information Processing Systems 26: the proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013 (NIPS 2013)*, pages 2184–2192, 2013.

Eric Brochu, Tyson Brochu, and Nando de Freitas. A Bayesian Interactive Optimization Approach to Procedural Animation Design. In *Proceedings of the 2010 Eurographics/ACM SIGGRAPH Symposium on Computer Animation, SCA 2010, Madrid, Spain, 2010*, pages 103–112, 2010.

Cameron Browne, Edward J. Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A Survey of Monte Carlo Tree Search Methods. *IEEE Trans. Comput. Intellig. and AI in Games*, 4(1):1–43, 2012.

Sébastien Bubeck. *Bandits Games and Clustering Foundations*. Theses, Université des Sciences et Technologie de Lille - Lille I, June 2010.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure Exploration in Multi-armed Bandits Problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT 2009)*, pages 23–37, 2009.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-Armed Bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

Sébastien Bubeck and Aleksandrs Slivkins. The Best of Both Worlds: Stochastic and Adversarial Bandits. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT 2012)*, pages 42.1–42.23, 2012.

Loc Bui, Ramesh Johari, and Shie Mannor. Committing Bandits. In *Advances in Neural Information Processing Systems 24: the proceedings of the 25th Annual Conference on Neural Information Processing Systems 2011 (NIPS 2011)*, pages 1557–1565, 2011.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.*, 41(3):1516–1541, 06 2013.

Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal Multi-Armed Bandits. In *Advances in Neural Information Processing Systems 21: the proceedings of the 22nd Annual Conference on Neural Information Processing Systems 2008 (NIPS 2008)*, pages 273–280, 2008.

Olivier Chapelle and Lihong Li. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems 24: the proceedings of the 25th Annual Conference on Neural Information Processing Systems 2011 (NIPS 2011)*, pages 2249–2257, 2011.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial Multi-Armed Bandit: General Framework and Applications. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 151–159, 2013.

Theodore Colton. A Model for Selecting One of Two Medical Treatments. *Journal of the American Statistical Association*, 58(302):pp. 388–400, 1963.

Richard Combes and Alexandre Proutiere. Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms. In *Proceedings of the 31th International Conference on Machine Learning (ICML 2014)*, pages 521–529, 2014.

Nick Craswell, Onno Zoeter, Michael J. Taylor, and Bill Ramsey. An experimental com-

parison of click position-bias models. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM 2008)*, pages 87–94, 2008.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of The 21st Conference on Learning Theory (COLT 2008)*, pages 355–366, 2008.

Yahel David and Nahum Shimkin. Infinitely Many-Armed Bandits with Unknown Value Distribution. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Proceedings, Part I*, pages 307–322, 2014.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Ann. Probab.*, 32(3):1902–1933, 07 2004.

Nikhil R. Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *ACM Conference on Electronic Commerce (EC 2009)*, pages 99–106, 2009.

Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005)*, pages 385–394, 2005.

Aurélien Garivier and Olivier Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT 2011)*, pages 359–376, 2011.

Nicola Gatti, Alessandro Lazaric, and Francesco Trovò. A truthful learning mechanism for multi-slot sponsored search auctions with externalities. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, pages 1325–1326, 2012.

Marco De Gemmis, Leo Iaquinta, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. Preference learning in recommender systems. In *Preference Learning (PL-09) ECML/PKDD-09 Workshop*, 2009.

J.C. Gittins and D.M. Jones. A Dynamic Allocation Index for the Sequential Design of Experiments. In J. Gani, editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, NL, 1974.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson Sampling for Complex Bandit Problems. In *Proceedings of the 31th International Conference on Machine Learning (ICML 2014)*, 2014.

Todd L. Graves and Tze Leung Lai. Asymptotically Efficient Adaptive Choice of Control Laws in Controlled Markov Chains. *SIAM Journal on Control and Optimization*, 35 (3):715–743, 1997.

Sudipto Guha and Kamesh Munagala. Multi-armed Bandits with Metric Switching Costs. In *Proceedings of the 36th Internatilonal Colloquium on Automata, Languages and Programming (ICALP 2009), Part II*, pages 496–507, 2009.

Stephen J Herschkorn, Erol Pekoez, and Sheldon M Ross. Policies without memory for the infinite-armed Bernoulli bandit under the average-reward criterion. *Probability in the Engineering and Informational Sciences*, 10(01):21–28, 1996.

Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Fidelity, Soundness, and Efficiency of Interleaved Comparison Methods. *Transactions on Information Systems*, 31(4):17:1–43, 2013.

Junya Honda and Akimichi Takemura. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *Proceedings of the 23rd Conference on Learning Theory (COLT 2010)*, pages 67–79, 2010.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

Junya Honda and Akimichi Takemura. Optimality of Thompson Sampling for Gaussian Bandits Depends on Priors. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS 2014)*, pages 375–383, 2014.

Senhua Huang, Xin Liu, and Zhi Ding. Opportunistic Spectrum Access in Cognitive Radio Networks. In *Proceedings of the 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2008)*, pages 1427–1435, 2008.

Patrick Hummel and R. Preston McAfee. Machine learning in an auction environment. In *Proceedings of the 23rd International World Wide Web Conference (WWW 2014)*, pages 7–18, 2014.

Emil Jerábek. Dual weak pigeonhole principle, Boolean complexity, and derandomization. *Ann. Pure Appl. Logic*, 129(1-3):1–37, 2004.

Thorsten Joachims. Evaluating Retrieval Performance Using Clickthrough Data. In *Text Mining*, pages 79–96. 2003.

Tackseung Jun. A survey on the bandit problem with switching costs. *De Economist*, 152 (4):513–541, 2004.

Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003)*, pages 583–588, 2003.

Emilie Kaufmann. *Analyse de stratágies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources.* Theses, TELECOM ParisTech, October 2014.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT 2012)*, pages 199–213, 2012.

David Kempe and Mohammad Mahdian. A Cascade Model for Externalities in Sponsored Search. In *Proceedings of the Internet and Network Economics, 4th International Workshop (WINE 2008)*, pages 585–596, 2008.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric

spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 681–690, 2008.

Tomás Kocák, Michal Valko, Rémi Munos, and Shipra Agrawal. Spectral Thompson Sampling. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI 2014)*, pages 1911–1917, 2014.

Levente Kocsis and Csaba Szepesvári. Bandit Based Monte-Carlo Planning. In *Proceedings of the 17th European Conference on Machine Learning (ECML 2006)*, pages 282–293, 2006.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem. In *Proceedings of The 28th Conference on Learning Theory (COLT 2015)*, pages 1141–1154, 2015a.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 1152–1161, 2015b.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret Lower Bound and Optimal Algorithm in Finite Stochastic Partial Monitoring. In *Advances in Neural Information Processing Systems 28: the proceedings of the 29th Annual Conference on Neural Information Processing Systems 2015 (NIPS 2015)*, 2015c.

Junpei Komiyama, Issei Sato, and Hiroshi Nakagawa. Multi-armed Bandit Problem with Lock-up Periods. In *Proceedings of the 5th Asian Conference on Machine Learning (ACML 2013)*, pages 116–132, 2013a.

Junpei Komiyama, Issei Sato, and Hiroshi Nakagawa. Multi-armed Bandit Problem with Lock-up Periods. *Transactions on Mathematical Modeling and its Applications (In Japanese)*, 6(3):11–22, 2013b.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson Sampling for 1-Dimensional Exponential Family Bandits. In *Advances in Neural Information Processing Systems 26: the proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013 (NIPS 2013)*, pages 1448–1456, 2013.

Tze Leung Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *Ann. Statist.*, 15(3):1091–1114, 09 1987.

Tze Leung Lai and Herbert Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

John Langford and Tong Zhang. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *Advances in Neural Information Processing Systems 20, Proceedings of the 21st Annual Conference on Neural Information Processing Systems 2007 (NIPS 2007)*, 2007.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th Inter-*

national Conference on World Wide Web (WWW 2010), pages 661–670, 2010.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz Bandits: Regret Lower Bound and Optimal Algorithms. In *Proceedings of The 27th Conference on Learning Theory (COLT 2014)*, pages 975–999, 2014.

Aditya Mahajan and Demosthenis Teneketzis. *Foundations and Applications of Sensor Management*, chapter Multi-Armed Bandit Problems, pages 121–151. Springer US, Boston, MA, 2008. ISBN 978-0-387-49819-5. URL `http://dx.doi.org/10.1007/978-0-387-49819-5_6`.

C.L Mallows and Herbert Robbins. Some problems of optimal sampling strategy. *Journal of Mathematical Analysis and Applications*, 8(1):90 – 103, 1964.

Microsoft Research. Microsoft Learning to Rank Datasets, 2010. URL `http://research.microsoft.com/en-us/projects/mslr/`.

Jonas Mockus. On Bayesian Methods for Seeking the Extremum. In *Optimization Techniques, IFIP Technical Conference, Novosibirsk, USSR, July 1-7, 1974*, pages 400–404, 1974.

Gergely Neu and Gábor Bartók. An Efficient Algorithm for Learning with Semi-bandit Feedback. In *Proceedings of the 24th International Conference on Algorithmic Learning Theory (ALT 2013)*, pages 234–248, 2013.

Pedro A. Ortega and Daniel A. Braun. A Minimum Relative Entropy Principle for Learning and Acting. *J. Artif. Int. Res.*, 38(1):475–511, May 2010. ISSN 1076-9757.

Antonio Piccolboni and Christian Schindelhauer. Discrete Prediction Games with Arbitrary Feedback and Loss. In *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 16-19, 2001, Proceedings*, pages 208–223, 2001.

Tao Qin, Wei Chen, and Tie-Yan Liu. Sponsored Search Auctions: Recent Advances and Future Directions. *ACM TIST*, 5(4):60:1–60:34, 2014.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr.*, 13(4):346–374, 2010.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the Twenty-Fifth International Conference (ICML 2008)*, pages 784–791, 2008a.

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, pages 43–52, 2008b.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the AMS*, 58:527–535, 1952.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly Parameterized Bandits. *Math. Oper. Res.*, 35(2):395–411, 2010.

Steven L. Scott. A modern Bayesian look at the multi-armed bandit. *Appl. Stoch. Model. Bus. Ind.*, 26(6):639–658, November 2010. ISSN 1524-1904.

Steven L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31:37–49, 2015. Special issue on actual impact and future perspectives on stochastic modelling in business and industry.

Yevgeny Seldin, Peter L. Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with Limited Advice and Multiarmed Bandits with Paid Observations. In *Proceedings of the 31th International Conference on Machine Learning (ICML 2014)*, pages 280–287, 2014.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems 25: the proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012 (NIPS 2012)*, pages 2960–2968, 2012.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

William R. Thompson. On The Likelihood That One Unknown Probability Exceeds Another In View Of The Evidence Of Two Samples. *Biometrika*, 25:285–294, 1933.

Taishi Uchiya, Atsuyoshi Nakamura, and Mineichi Kudo. Algorithms for Adversarial Bandit Problems with Multiple Plays. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory (ALT 2010)*, pages 375–389, 2010.

Tanguy Urvoy, Fabrice Clérot, Raphaël Feraud, and Sami Naamane. Generic Exploration and K-armed Voting Bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 91–99, 2013.

Hamed Valizadegan, Rong Jin, and Shijun Wang. Learning to trade off between exploration and exploitation in multiclass bandit prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2011)*, pages 204–212, 2011.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for Infinitely Many-Armed Bandits. In *Advances in Neural Information Processing Systems 21, Proceedings of the 22nd Annual Conference on Neural Information Processing Systems 2008 (NIPS 2008)*, pages 1729–1736, 2008.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient Learning in Large-Scale Combinatorial Semi-Bandits. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 1113–1122, 2015.

Min Xu, Tao Qin, and Tie-Yan Liu. Estimation Bias in Multi-Armed Bandit Algorithms for Search Advertising. In *Advances in Neural Information Processing Systems 26: the proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013 (NIPS 2013)*, pages 2400–2408. 2013.

Jia Yuan Yu and Shie Mannor. Unimodal Bandits. In *Proceedings of the 28th International*

Conference on Machine Learning (ICML 2011), pages 41–48, 2011.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed Dueling Bandits Problem. In *Proceedings of the 22nd Conference on Learning Theory (COLT 2009)*, 2009.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556, 2012.

Yisong Yue and Thorsten Joachims. Beat the Mean Bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 241–248, 2011.

Omar Zaidan and Chris Callison-Burch. Crowdsourcing Translation: Professional Quality from Non-Professionals. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1220–1229, 2011.

Masrour Zoghi, Shimon Whiteson, and Maarten de Rijke. MergeRUCB: A method for large-scale online ranker evaluation. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM 2015)*, 2015.

Masrour Zoghi, Shimon Whiteson, Maarten de Rijke, and Rémi Munos. Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM 2014)*, pages 73–82, 2014a.

Masrour Zoghi, Shimon Whiteson, Rémi Munos, and Maarten de Rijke. Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem. *CoRR*, abs/1312.3393v2, 2013. URL `http://arxiv.org/abs/1312.3393v2`.

Masrour Zoghi, Shimon Whiteson, Rémi Munos, and Maarten de Rijke. Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem. In *Proceedings of the 31th International Conference on Machine Learning (ICML 2014)*, pages 10–18, 2014b.