Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays

Junpei Komiyama JUNPEI@KOMIYAMA.INFO
Junya Honda HONDA@STAT.T.U-TOKYO.AC.JP
Hiroshi Nakagawa NAKAGAWA@DL.ITC.U-TOKYO.AC.JP
The University of Tokyo, Japan

Abstract

We discuss a multiple-play multi-armed bandit (MAB) problem in which several arms are selected at each round. Recently, Thompson sampling (TS), a randomized algorithm with a Bayesian spirit, has attracted much attention for its empirically excellent performance, and it is revealed to have an optimal regret bound in the standard single-play MAB problem. In this paper, we propose the multiple-play Thompson sampling (MP-TS) algorithm, an extension of TS to the multiple-play MAB problem, and discuss its regret analysis. We prove that MP-TS for binary rewards has the optimal regret upper bound that matches the regret lower bound provided by Anantharam et al. (1987). Therefore, MP-TS is the first computationally efficient algorithm with optimal regret. 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. Introduction

The multi-armed bandit (MAB) problem is one of the most well-known instances of sequential decision-making problems in uncertain environments, which can model many real-world scenarios. The problem involves conceptual entities called arms. At each round, the forecaster draws one of K arms and receives a corresponding reward. The aim of the forecaster is to maximize the cumulative reward over rounds, and the forecaster's performance is usually measured by a regret, which is the gap between his or her cumulative reward and that of an optimal drawing policy. Throughout the rounds, the forecaster faces an "exploration vs. exploitation" dilemma. On one hand, the forecaster wants to exploit the information that he or she has gathered up to the previous round by selecting seemingly good arms. On the other hand, there is always a possibility that the other arms have been underestimated, which motivates him or her to explore seemingly bad arms in order to gather their information. To resolve this dilemma, the forecaster uses an algorithm to control the number of draws for each arm.

In the stochastic MAB problem, which is the most widely studied version of the MAB problem, it is assumed that each arm is associated with a distinct probability distribution. While there have been many theoretical studies on the infinite setting in which future rewards are geometrically discounted (e.g., the Gittins index (Gittins & Jones, 1974)), recent availability of massive data has led to a finite horizon setting in which every reward has the same importance. In this work, we focus on the latter setting.

There has been significant progress in this setting of the MAB problem. In particular, the upper confidence bound (UCB) algorithm (Auer et al., 2002) has been widely used and studied for its computational simplicity and customizability. Whereas the coefficient of the leading logarithmic term in UCB is larger than the theoretical lower bound given by Lai & Robbins (1985), algorithms have been proposed that achieve this bound, such as DMED (Honda & Takemura, 2010), $\mathcal{K}_{\inf}$, and KL-UCB (Cappé et al., 2013).

Moreover, Thompson sampling (TS) (Thompson, 1933) has recently attracted attention for its excellent performance (Scott, 2010; Chapelle & Li, 2011) and it has been revealed to be applicable to even a wider class of problems (Agrawal & Goyal, 2013a; Russo & Roy, 2013; Osband et al., 2013; Kocák et al., 2014; Guha & Munagala, 2014). Thompson sampling is an old heuristic that has a spirit of Bayesian inference and selects an arm based on posterior samples of the expectation of each arm. It has been shown that TS has an optimal regret bound (Agrawal & Goyal,

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

2012; Kaufmann et al., 2012; Agrawal & Goyal, 2013b).

1.1. Multiple-play MAB problem

The literature mentioned above has specifically dealt with the MAB problem in which a single arm is selected and drawn at each round. Let us call this problem single-play MAB (SP-MAB). While the SP-MAB 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 MAB problem in which several arms can be selected multiple-play MAB (MP-MAB). Examples of the situations that can be modeled as an MP-MAB problem include the followings.

- **Example 1 (placement of online advertisements):** a web site 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 web sites can select to display. 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 web sites 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-MAB problem in which channels and the permission of communication are arms and rewards, respectively.

There have been several studies on the MP-MAB problem. Anantharam et al. (1987) derived an asymptotic lower bound on the regret for this problem and proposed an algorithm to achieve this bound. Because their algorithm requires certain statistics that are difficult to compute, efficiently computable MP-MAB 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 constant factors of these regret bounds do not match the lower bound. Therefore, it is unknown whether the optimal regret bound for the MP-MAB problem is achievable by using a computationally efficient algorithm.

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

Contribution: Our contributions are as follows.

- **TS-based algorithm for the MP-MAB problem and its optimal regret bound:** the first and main contribution of this paper is an extension of TS to the multiple play case, which we call MP-TS. We prove that MP-TS for binary rewards achieves an optimal regret bound. To the best of our knowledge, this paper is the first to provide a computationally efficient algorithm in the MP-MAB problem with the optimal regret bound by Anantharam et al. (1987).
- **Novel analysis technique:** to solve the difficulty in the combinatorial structure of the MP-MAB 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-MAB 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 analyses on the regret structure of the MP-MAB problem. We confirm that IMP-TS improves the performance of MP-TS without increasing computational complexity.

2. Problem Setup

Let there be K arms. Each arm $i \in [K] = \{1, 2, \ldots, K\}$ is associated with a probability distribution $\nu_i = \text{Bernoulli}(\mu_i)$, $\mu_i \in (0, 1)$. 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 $X_i(t)$ of each selected arm i is i.i.d. samples from ν_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 $\hat{\mu}_i(t)$ 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 Appendix A.1, which is in Supplementary Material. 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

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

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

3. Regret Bounds

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

3.1. Regret bound for SP-MAB problem

The SP-MAB problem, which has been thoroughly studied in the fields of statistics and machine learning, is a special case of the MP-MAB problem with $L = 1$. The optimal regret bound in the SP-MAB problem was given by Lai & 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

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

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

$$\mathbb{E}[\mathrm{Reg}(T)] = \sum_{i \neq 1} N_i(T+1) \Delta_{i,1}, \tag{2}$$

where $\Delta_{i,j} = \mu_j - \mu_i$. Therefore, inequality (1) directly leads to the regret lower bound

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \sum_{i \neq 1} \left(\frac{(1 - o(1)) \Delta_{i,1}}{d(\mu_i, \mu_1)} \right) \log T. \tag{3}$$

One may think that applying the techniques of the SP-MAB problem would directly yield an optimal bound for a more general MP-MAB problem. However, this is not the case. In short, the difficulty in analyzing the regret on the MP-MAB 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-MAB problem in which L is not restricted to one.

A MP-MAB instance with K=4, L=2

μ_1=0.10, μ_2=0.09 ⇝ optimal arms

μ_3=0.08, μ_4=0.07 ⇝ suboptimal arms

	Game 1	Game 2
t=1	$I(1) = \{1, 2\}$	$I(1) = \{1, 3\}$
	$(r(1) = 0)$	$(r(1)=0.01)$
t=2	$I(2) = \{3, 4\}$	$I(2) = \{1, 4\}$
	$(r(2) = 0.04)$	$(r(2)=0.02)$
	Regret(2)=0.04	Regret(2)=0.03

Figure 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.

3.2. Extension to MP-MAB problem

The regret lower bound in the MP-MAB problem, which is the generalization of inequality (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

$$\mathbb{E}[N_i(T+1)] \geq \left(\frac{1 - o(1)}{d(\mu_i, \mu_L)} \right) \log T. \tag{4}$$

Unlike in the SP-MAB problem, the regret in the MP-MAB problem is not uniquely determined by the number of suboptimal arm draws. As illustrated in Figure 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

$$\begin{aligned} \sum_{j \in [L]} \mu_j - \sum_{i \in I(t)} \mu_i &= \sum_{j \in [L] \setminus I(t)} \mu_j - \sum_{i \in I(t) \setminus [L]} \mu_i \\ &\geq \sum_{i \in I(t) \setminus [L]} (\mu_L - \mu_i), \end{aligned} \tag{5}$$

Algorithm 1 Multiple-play Thompson sampling (MP-TS) for binary rewards

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

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

$$\begin{aligned}\mathrm{Reg}(T) &\geq \sum_{t=1}^{T} \sum_{i \in I(t) \setminus [L]} (\mu_L - \mu_i) \\ &= \sum_{i \in [K] \setminus [L]} (\mu_L - \mu_i) N_i(T+1) \qquad (6)\end{aligned}$$

which, combined with (4), leads to the regret lower bound by Anantharam et al. (1987) that any strongly consistent algorithm satisfies

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \sum_{i \in [K] \setminus [L]} \frac{(1 - o(1))\Delta_{i,L}}{d(\mu_i, \mu_L)} \log T. \qquad (7)$$

3.3. Necessary condition for an optimal algorithm

In Sections 3.1 and 3.2, we saw that the derivations of the regret bounds are analogous between the SP-MAB and MP-MAB 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 (2) in the SP-MAB problem, whereas it is given as inequality (6) in the MP-MAB problem. This means that, an algorithm achieving the asymptotic lower bound (4) on $N_i(T)$ does not always achieve the asymptotic regret bound (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 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-MAB 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 (4) on the number of draws of suboptimal arms. Whereas this does not lead to the 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-MAB. 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.

4. Multiple-play Thompson Sampling Algorithm

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

In Section 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-MAB problem.

Note that it is easy to extend other asymptotically optimal SP-MAB algorithms, such as KL-UCB, to the MP-MAB problem. Nevertheless, we were not able to prove the optimality of these algorithms for the MP-MAB problem though the achievability of the bound (4) on $N_i(T)$ is easily proved, and the simulation results in Section 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 (or other asymptotically optimal SP-MAB algorithms) can have the opti-

mal regret in the MP-MAB 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 $\theta_i(t)$ 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 policies), 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. Optimal Regret Bound

In this section, we state the main theoretical result (Theorem 1). The analysis that leads to this theorem is discussed in Section 6.

Theorem 1. (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 $O(\log\log T) = o(\log T)$.

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

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i\in[K]\setminus[L]} \frac{\Delta_{i,L}\log T}{d(\mu_i,\mu_L)} + O((\log T)^{2/3}) \quad (8)$$

and we see that MP-TS achieves the asymptotic bound in (7).

Expected regret and high-probability regret: Anantharam et al. (1987) originally derived a regret lower bound in a stronger form than (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 (8) we can easily see that MP-TS satisfies

$$\lim_{T\to\infty} \Pr\left[\frac{\mathrm{Reg}(T)}{\log T} \leq \sum_{i\in[K]\setminus[L]} \frac{(1+\epsilon)\Delta_{i,L}}{d(\mu_i,\mu_L)}\right] = 1, \quad (9)$$

that is, MP-TS is also asymptotically optimal in the sense of high probability. Since an algorithm satisfying (9) is not always optimal in the sense of expectation, our result, the expected optimal regret bound, is also stronger in this sense than the high-probability bound by Gopalan et al. (2014).

6. Regret Analysis

We first define some additional notation that are useful for our analysis in Section 6.1 then analyze the regret bound in Section 6.2. The proofs of all the lemmas, except for Lemma 2, are given in the Appendix.

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: Now, 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 $\theta^*(t) = \max_{i\in[K]}^{(L)} \theta_i(t)$ as the L-th largest posterior sample at round t (i.e., the minimum posterior sample among the selected arms), and $\theta^{**}_{\setminus i,j}(t) = \max_{k\in[K]\setminus\{i,j\}}^{(L-1)} \theta_k(t)$ 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.

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

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.

6.2. Proof of Theorem 1

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

$$\Delta_i(t) = \begin{cases} (\max_{j\in[L]\setminus I(t)} \mu_j) - \mu_i & \text{if } I(t) \neq [L], \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

and

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

From inequality (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 1. 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-MAB problem, mostly from Agrawal & Goyal (2013b), and some from Honda & Takemura (2014).

Lemma 2. *The regret by drawing suboptimal arm $i > L$ is decomposed as:*

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

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-MAB 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 2. The contribution of suboptimal arm i to the regret is decomposed as follows. By using the fact $\Delta_i(t) \leq 1$ and the following decomposition of an event

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

we have

$$\begin{aligned}\mathrm{Reg}_i(T) &= \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t)\}\Delta_i(t) \\ &\leq \sum_{t=1}^{T} \mathbf{1}\{\mathcal{B}^c(t)\} + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i^c(t)\} \\ &\quad + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{B}(t), \mathcal{D}_i^c(t)\} \\ &\quad + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t)\}\Delta_i(t). \end{aligned} \tag{11}$$

Recall that $\Delta_i(t)$ is defined as (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,

$$\begin{aligned}&\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t)\}\Delta_i(t) \\ &\leq \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t)\}\Delta_{i,L} \\ &\quad + \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \bigcup_{j\in[K]\setminus([L-1]\cup\{i\})} \mathcal{A}_j(t)\} \\ &\leq \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{D}_i(t)\}\Delta_{i,L} \\ &\quad + \sum_{j\in[K]\setminus([L-1]\cup\{i\})} \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \mathcal{A}_j(t)\} \\ &\leq N_i^{\mathrm{suf}}(T)\Delta_{i,L} \\ &\quad + \sum_{j\in[K]\setminus([L-1]\cup\{i\})} \sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t), \mathcal{C}_i(t), \mathcal{D}_i(t), \mathcal{A}_j(t)\}. \end{aligned} \tag{12}$$

Summarizing (11) and (12) completes the proof. □

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

Lemma 3. (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:

$$\mathbb{E}[(\mathrm{A})] = O\left(\frac{1}{(\mu_L - \mu_L^{(-)})^2}\right) = O\left(\frac{1}{\delta^2}\right), \tag{13}$$

$$\mathbb{E}[(\mathrm{B})] = O(1), \tag{14}$$

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

and (15)

$$\mathbb{E}[(\mathrm{D})] \leq 2 + \frac{1}{d(\mu_i^{(+)},\mu_i)} = O\left(\frac{1}{\delta^2}\right). \tag{16}$$

The proof of Lemma 3 is in Appendix A.4. Lemma 3 states that terms (A), (B), and (D) are $O(1/\delta^2)$. Moreover, the following lemma bounds term (C).

Lemma 4. (Asymptotic convergence 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 4 is in Appendix A.5. 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

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{i\in[K]\setminus[L]} \mathbb{E}[\mathrm{Reg}_i(T)] \leq \sum_{i\in[K]\setminus[L]} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{A}_i(t)\}\Delta_i(t)\right]$$

$$\leq \sum_{i\in[K]\setminus[L]} \left\{\mathbb{E}\left[(\mathrm{A})+(\mathrm{B})+(\mathrm{C})+(\mathrm{D})\right] + N_i^{\mathrm{suf}}(T)\Delta_{i,L}\right\}$$

$$\leq \underbrace{\sum_{i\in[K]\setminus[L]} \frac{\Delta_{i,L}\log T}{(1-c_i\delta)d(\mu_i,\mu_L)}}_{\text{main term}} + \underbrace{O\left(\frac{1}{\delta^2}\right)}_{C_a} + \underbrace{O(\log\log T)}_{C_b}.$$

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

7. Experiment

We ran a series of computer simulations[1] 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 & 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-MAB problem[2]. 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-MAB problem at each round draws the top-L arms of the UCB indices $\hat{\mu}_i + \sqrt{(3\log t)/(2N_i(t))}$. MP-KL-UCB is the algorithm that selects the top-L arms in accordance with the KL-UCB index $\sup_{q\in[\hat{\mu}_i(t),1]}\{q | N_i(t)d(\hat{\mu}_i(t),q) \leq \log t\}$.

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 $L = 3$.

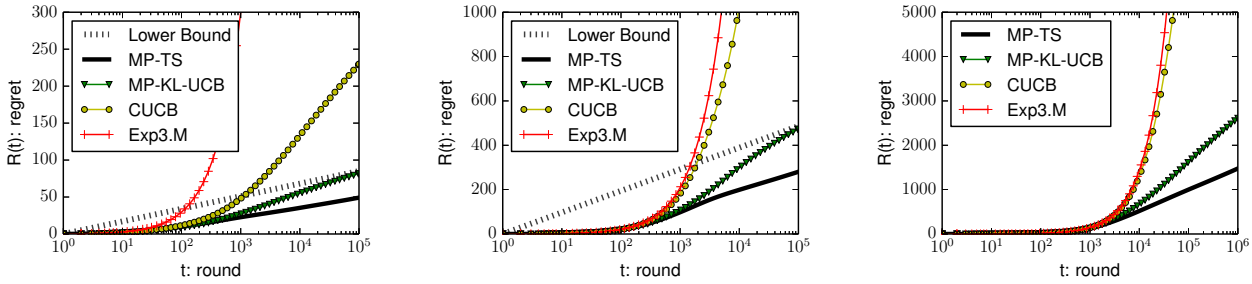
Scenario 3 (many-armed bandits, online advertisement based CTRs): we conducted another set of experiments with arms whose expectations were based on the dataset provided for KDD Cup[3] 2012 track 2. The dataset involves a click log on soso.com (a large-scale search engine serviced by Tencent), which 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 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 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

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

[2]Note that, Exp3.M is designed for the adversarial setting in which the rewards of arms are not necessarily stationary.

[3]https://www.kddcup2012.org/



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Figure 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 (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).

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.

7.1. Improvement of MP-TS based on the empirical means

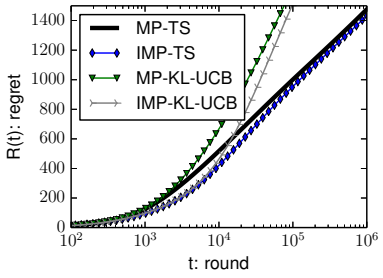


Figure 3. Before/after comparison of MP-TS. All settings (except for algorithms) are the same as that of Scenario 3.

We now introduce an improved version of MP-TS (IMP-TS). In the theoretical analysis of the MP-MAB 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 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.

8. Discussion

We extended TS to the multiple-play setting and proved its 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 in this paper 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-MAB 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 & Mahdian, 2008; Aggarwal et al., 2008) that corrects position-dependent bias in Appendix A.2.
- **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 & Katehakis, 1996; Honda & Takemura, 2014) is interesting.

Acknowledgements

We gratefully acknowledge the insightful advice from Issei Sato and Tao Qin. We thank the anonymous reviewers for their useful comments. This work was supported in part by JSPS KAKENHI Grant Number 26106506.

References

Aggarwal, Gagan, Feldman, Jon, Muthukrishnan, S., and Pál, Martin. Sponsored search auctions with markovian users. In *WINE*, pp. 621–628, 2008.

Agrawal, Shipra and Goyal, Navin. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pp. 39.1–39.26, 2012.

Agrawal, Shipra and Goyal, Navin. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pp. 127–135, 2013a.

Agrawal, Shipra and Goyal, Navin. Further optimal regret bounds for thompson sampling. In *AISTATS*, pp. 99–107, 2013b.

Anantharam, V., Varaiya, P., and Walrand, J. 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.

Auer, Peter, Cesa-bianchi, Nicoló, and Fischer, Paul. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47:235–256, 2002.

Burnetas, A.N. and Katehakis, M.N. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Cappé, Olivier, Garivier, Aurélien, Maillard, Odalric-Ambrym, Munos, Rémi, and Stoltz, Gilles. Kullback-leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.*, 41(3):1516–1541, 06 2013.

Chapelle, Olivier and Li, Lihong. An empirical evaluation of thompson sampling. In *NIPS*, pp. 2249–2257, 2011.

Chen, Wei, Wang, Yajun, and Yuan, Yang. Combinatorial multi-armed bandit: General framework and applications. In *ICML*, pp. 151–159, 2013.

Craswell, Nick, Zoeter, Onno, Taylor, Michael J., and Ramsey, Bill. An experimental comparison of click position-bias models. In *WSDM*, pp. 87–94, 2008.

Garivier, Aurélien and Cappé, Olivier. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, pp. 359–376, 2011.

Gatti, Nicola, Lazaric, Alessandro, and Trovò, Francesco. A truthful learning mechanism for multi-slot sponsored search auctions with externalities. In *AAMAS*, pp. 1325–1326, 2012.

Gittins, J.C. and Jones, D.M. A dynamic allocation index for the sequential design of experiments. In Gani, J. (ed.), *Progress in Statistics*, pp. 241–266. North-Holland, Amsterdam, NL, 1974.

Gopalan, Aditya, Mannor, Shie, and Mansour, Yishay. Thompson sampling for complex bandit problems. In *ICML*, 2014.

Guha, Sudipto and Munagala, Kamesh. Stochastic regret minimization via thompson sampling. In *COLT*, pp. 317–338, 2014.

Honda, Junya and Takemura, Akimichi. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *COLT*, pp. 67–79, 2010.

Honda, Junya and Takemura, Akimichi. Optimality of thompson sampling for gaussian bandits depends on priors. In *AISTATS*, pp. 375–383, 2014.

Huang, Senhua, Liu, Xin, and Ding, Zhi. Opportunistic spectrum access in cognitive radio networks. In *INFOCOM*, pp. 1427–1435, 2008.

Kaufmann, Emilie, Korda, Nathaniel, and Munos, Rémi. Thompson sampling: An asymptotically optimal finite-time analysis. In *ALT*, pp. 199–213, 2012.

Kempe, David and Mahdian, Mohammad. A cascade model for externalities in sponsored search. In *WINE*, pp. 585–596, 2008.

Kocák, Tomás, Valko, Michal, Munos, Rémi, and Agrawal, Shipra. Spectral thompson sampling. In *AAAI*, pp. 1911–1917, 2014.

Korda, Nathaniel, Kaufmann, Emilie, and Munos, Rémi. Thompson sampling for 1-dimensional exponential family bandits. In *NIPS*, pp. 1448–1456, 2013.

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

Osband, Ian, Russo, Daniel, and Roy, Benjamin Van. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, pp. 3003–3011, 2013.

Russo, Daniel and Roy, Benjamin Van. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pp. 2256–2264, 2013.

Scott, Steven L. A modern bayesian look at the multi-armed bandit. *Appl. Stoch. Model. Bus. Ind.*, 26(6):639–658, November 2010. ISSN 1524-1904.

Thompson, William R. On The Likelihood That One Unknown Probability Exceeds Another In View Of The Evidence Of Two Samples. *Biometrika*, 25:285–294, 1933.

Uchiya, Taishi, Nakamura, Atsuyoshi, and Kudo, Mineichi. Algorithms for adversarial bandit problems with multiple plays. In *ALT*, pp. 375–389, 2010.