Thompson Sampling for Multi-Armed Bandit and Complex Online Problems

Harsh Gupta
CS598NJ Course Project
hgupta10@illinois.edu

Abstract

In this project report, we present the analysis of Thompson Sampling (TS), a popular Bayesian algorithm for the standard stochastic multi-armed bandit problem. Thompson sampling had little to no theoretical guarantees until very recently, although it was known to perform well in practice. We primarily focus on the analysis by Agrawal and Goyal [2013] that obtains an optimal problem-dependent upper bound and a near-optimal problem-independent upper bound for the expected regret of Thompson sampling, when applied to the standard stochastic multi-armed bandit problem. We also present a short discussion on more complex online sequential decision making problems and the use of Thompson sampling in their context.

1 Introduction

The stochastic multi-armed bandit (MAB) problem has been a topic of detailed study over the last two decades (see Bubeck et al. [2012] for a survey). The standard stochastic MAB problem can be described as follows: At any time t, there are K arms available to a player to choose from. Each arm i has an underlying Bernoulli distribution from which its reward $r_i(t) \in \{0, 1\}$ is drawn at time t. The rewards are assumed to be drawn i.i.d. across different time slots. Let μ_i be the expected reward of arm i. Without loss of generality, let us assume that arm 1 is the optimal arm, i.e., $\mu^* = \mu_1 > \mu_i, \forall i \neq 1$. Also, let $\Delta_i = \mu^* - \mu_i$. If the player chooses arm i at time t, it receives the reward $r_i(t)$. The objective of the stochastic MAB problem is to devise an efficient algorithm, which takes the history of actions played and rewards received for the first $t-1$ time slots, and then devises a strategy (deterministic or stochastic) to select an arm to play at time t. To quantify the performance of any MAB algorithm, there are different metrics. One obvious metric is the expected reward. Let $i(t)$ denote the arm played by the algorithm at time t, then the expected reward until time T is defined as:

$$\mathbb{E}[\Lambda(T)] = \sum_{t=1}^{T} \mathbb{E}[\mu_{i(t)}], \tag{1}$$

where the expectation is with respect to the randomness in the choice of action $i(t)$ (which includes the randomness due to the stochastic rewards). A more popular and commonly used metric is the expected regret, which essentially quantifies the difference in the performance of a MAB algorithm from an algorithm that simply pulls the optimal arm at each time slot. Formally, the expected regret is defined as:

$$\mathbb{E}[R(T)] = \sum_{t=1}^{T} \mathbb{E}[\mu^* - \mu_{i(t)}] = \sum_{i \neq 1} \mathbb{E}[N_i(T+1)]\Delta_i, \tag{2}$$

where $N_i(T+1)$ is the number of times arm i has been played until the end of T time slots. Again, the expectation in the above equation is with respect to the randomness in the choice of the action $i(t)$.

Several popular algorithms for the standard MAB problem exist in literature (see Bubeck et al. [2012] for more information). They can be broadly classified into two classes: Frequentist and Bayesian. The frequentist class of methods includes the popular UCB algorithm (see Auer et al. [2002]) and its various variants. On the other hand, Thompson sampling, introduced by William Thompson in 1933 (see Thompson [1933]), is one of the most popular Bayesian algorithms. Recently, it attracted a lot of attention when it was shown to perform well in practice for the MAB problem (see Chapelle and Li [2011]). This led to a growing theoretical interest in proving guarantees for it. We will now discuss what the Thompson sampling algorithm is, before we move on to its analysis in the next section.

1.1 Thompson Sampling Algorithm

Let ψ denote the set of possible expected reward vectors, such that the true expected reward vector belongs to the set, i.e., $(\mu_1, \mu_2, ..., \mu_K) = \bar{\mu} \in \psi$. The basic idea behind the Thompson sampling algorithm is the following:

1. At any time t, start with a prior distribution $\mathbb{P}_t(\theta)$ to sample a mean reward vector $\theta = (\theta_1, \theta_2, ..., \theta_K)$ from the set ψ.
2. Sample $\theta(t) \sim \mathbb{P}_t$.
3. Play the optimal arm for $\theta(t)$, i.e., play $i(t) = \arg\max_i \theta_i(t)$.
4. Receive the reward $r_{i(t)}(t)$.
5. Posterior update: $\mathbb{P}_{t+1}(\theta) \propto \mathbb{P}(r_{i(t)}(t)|\theta)\mathbb{P}_t(\theta)$.

In this report, we focus exclusively on the case when the reward distribution for each arm is Bernoulli. A natural prior distribution for the Bernoulli distribution parameters is the Beta distribution, since the Beta distribution is the conjugate prior of the Bernoulli distribution. The p.d.f. for a Beta(α, β) distribution is given as:

$$p_{a,b}(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, x \in [0,1] \tag{3}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function and $\Gamma(x) = \int_0^\infty z^{x-1}e^{-z}dz$ is the Gamma function.

The advantage of using the Beta distribution as a prior is that it's posterior update step for a Bernoulli trial is very straightforward. If a success is obtained in a Bernoulli trial, whose true mean parameter is being estimated by Beta(α, β), the posterior update simply gives the Beta$(\alpha+1, \beta)$ distribution. Similarly, if a failure is obtained in the Bernoulli trial, the posterior update simply gives the Beta$(\alpha, \beta+1)$ distribution. We present Thompson sampling using Beta priors for the MAB problem as Algorithm 1.

Remark 1. *Although we focus on the MAB problem with Bernoulli rewards in this report, handling more general distributions on $[0,1]$ is relatively straightforward (see Agrawal and Goyal [2012]). Since there is only a minor technical difference between the two cases, we stick to Bernoulli rewards to make the analysis more insightful.*

Remark 2. *One can also use different priors for the Thompson sampling algorithm. Although we use Beta priors, Agrawal and Goyal [2013] also analyze Thompson sampling using Gaussian priors. Again, the difference in the two cases doesn't give rise to additional technical insights, therefore, we stick to Thompson sampling using Beta priors in this report.*

2 Analysis of Thompson Sampling with Beta Priors

In this section, we present the analysis of Algorithm 1. More precisely, we present upper bounds on the expected regret. First, let us start with stating the results.

Theorem 1. *For the K-armed standard stochastic multi-armed bandit problem, Algorithm 1 has the following problem-dependent expected regret, for any $\epsilon > 0$:*

$$\mathbb{E}[R(T)] \leq (1+\epsilon)\sum_{i=2}^{K} \frac{\ln T}{d(\mu_i, \mu_1)} + O(\frac{K}{\epsilon^2}),$$

in T time slots, where $d(\mu_i, \mu_1)$ is the KL-divergence between Bernoulli(μ_i) and Bernoulli(μ_1) random variables. The big-Oh notation assumes $\mu_i, \Delta_i, \forall i = 1, 2, ..., K$ to be constants.

Algorithm 1 Thompson sampling algorithm with Beta priors

for each arm $i = 1, 2, ..., K$, set $S_i = 0$ and $F_i = 0$.
for each $t = 1, 2, \ldots$:

1. For all arms i, draw $\theta_i(t) \sim \text{Beta}(S_i + 1, F_i + 1)$.
2. Play arm $i(t)$, where $i(t) = \arg\max_i \theta_i(t)$.
3. Receive the reward $r_{i(t)}(t)$.
4. (Posterior Update for Prior) If $r_{i(t)}(t) = 1$, set $S_{i(t)} = S_{i(t)} + 1$. Else if $r_{i(t)}(t) = 0$, set $F_{i(t)} = F_{i(t)} + 1$.

end for

Theorem 2. *For the K-armed standard stochastic multi-armed bandit problem, Algorithm 1 has the following problem-independent expected regret:*

$$\mathbb{E}[R(T)] \leq O(\sqrt{KT \ln T}),$$

in T time slots. The big-Oh notation only hides the absolute constants.

In order to prove the above bounds, we will first upper bound the number of times we pick any sub-optimal action i ($i \neq 1$) until time T. Eventually, to obtain the upper bound on total expected regret, we will simply sum the regret until time T due to each sub-optimal arm. In order to proceed with the analysis, we will lay down some definitions. Since we reproduce the analysis in Agrawal and Goyal [2013] to make it more lucid, some definitions and notation might be slightly different from theirs.

Definition 1. ***(Parameters** $N_i(t), i(t), S_i(t)$ **and** $\hat{\mu}_i(t)$**).** *Let $i(t)$ be the arm played by the Thompson sampling algorithm at time slot t. Let $N_i(t)$ denote the number of times arm i has been played until the end of time slot $t-1$. Let $S_i(t)$ denote the number of times arm i gave a reward of 1 among all its plays until the end of time slot $t-1$. Moreover, $\hat{\mu}_i(t)$ is defined as the empirical mean of the reward outcomes for arm i until the end of time slot $t-1$, i.e., $\hat{\mu}_i(t) = \frac{\sum_{j=1:i(j)=i}^{t-1} r_i(j)}{N_i(t)+1}$.*

Definition 2. ***(Thresholds** x_i, y_i**)** *For each arm i ($i \neq 1$), we will choose two thresholds x_i and y_i such that $\mu_i < x_i < y_i < \mu_1$. The choice of exact values of x_i and y_i will be presented in the proof.*

Definition 3. ***(Events** $E_i^{\mu}(t), E_i^{\theta}(t)$**)** *We define the event $E_i^{\mu}(t)$ as the event such that $\hat{\mu}_i(t) \leq x_i$. Similarly, $E_i^{\theta}(t)$ is the event such that $\theta_i(t) \leq y_i$.*

$E_i^{\mu}(t)$ defines the event that the empirical average of the reward outcomes of arm i (until time $t-1$) does not deviate too much from the true expected value μ_i. Similarly, $E_i^{\theta}(t)$ defines the event that the sampled expected reward for the arm i (by TS prior distribution at time t) does not deviate too much from μ_i. These are good events, which we will show hold with high probability.

Definition 4. ***(Filtration** $\mathscr{F}_{t-1}$**)** *We define the filtration $\mathscr{F}_{t-1}$ as the history of arms played and their reward outcomes until the end of time slot $t-1$, i.e., $\mathscr{F}_{t-1} = \{i(j), r_{i(j)}(j); j = 1, ..., t-1\}$.*

Definition 5. ***(Parameters** τ_i **and** $p_{i,t}$**).** *Let τ_i denote the time when the optimal arm 1 is transmitted the i^{th} time (for $i \geq 1$). Also, let $\tau_0 = 0$. We define the probability $p_{i,t}$ as, $p_{i,t} = \mathbb{P}(\theta_1(t) > y_i | \mathscr{F}_{t-1})$.*

A point worth noting is that, for every arm i, $\mathscr{F}_{t-1}$ determines $p_{i,t}$, $S_i(t), N_i(t), \hat{\mu}_i(t)$, the distribution of $\theta_i(t)$ and whether the event $E_i^{\mu}(t)$ is true or not.To bound the expected number of times we play arm i, we split the expectation into three different terms based on the occurrences of the events $E_i^{\mu}(t)$ and $E_i^{\theta}(t)$:

$$\begin{aligned}\mathbb{E}[N_i(T+1)] &= \sum_{t=1}^{T} \mathbb{P}(i(t) = i) \\ &= \sum_{t=1}^{T} \mathbb{P}(i(t) = i, E_i^{\theta}(t), E_i^{\mu}(t)) + \sum_{t=1}^{T} \mathbb{P}(i(t) = i, \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)) + \sum_{t=1}^{T} \mathbb{P}(i(t) = i, \overline{E_i^{\mu}(t)})\end{aligned} \tag{4}$$

where $\bar{A}$ denotes the complement of event A.

The intuition behind the above split is the following: $E_i^\theta(t)$ and $E_i^\mu(t)$ are good events, and as mentioned earlier, we will show they hold with high probability. Under these good events, we will show that the probability of selecting arm i is upper bounded by a linear function of the probability of selecting the optimal arm 1. Therefore, the overall contribution of this term in the plays of arm i is low. The other two terms correspond to low probability events, so they are naturally small. Therefore, we can bound the number of times arm i is played to a small quantity.

We will consider the three terms separately and derive an upper bound for each of them. Eventually we will add the upper bounds to get the upper bound on $\mathbb{E}[N_i(T+1)]$. To this end, we present the first lemma showing that the probability of selecting a sub-optimal arm i at any time slot is upper bounded by a linear function of the probability of selecting the optimal arm 1:

Lemma 1. *For all $t \in [1, T]$, and $i \neq 1$, we have:*

$$\mathbb{P}(i(t) = i, E_i^\mu(t), E_i^\theta(t) | \mathscr{F}_{t-1}) \leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}(i(t) = 1, E_i^\mu(t), E_i^\theta(t) | \mathscr{F}_{t-1}).$$

Proof. Since $\mathscr{F}_{t-1}$ determines the status of the event $E_i^\mu(t)$, we assume that the event took place as otherwise the LHS of the result is 0 and hence the lemma holds trivially. Therefore, we just need to show the following:

$$\mathbb{P}(i(t) = i | \mathscr{F}_{t-1}, E_i^\theta(t)) \leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}(i(t) = 1 | \mathscr{F}_{t-1}, E_i^\theta(t)). \tag{5}$$

For any sub-optimal arm i, we have:

$$\begin{aligned} \mathbb{P}(i(t) = i | \mathscr{F}_{t-1}, E_i^\theta(t)) &\leq \mathbb{P}(\theta_j(t) \leq y_i, \forall j | \mathscr{F}_{t-1}, E_i^\theta(t)) \\ &= \mathbb{P}(\theta_1(t) \leq y_i | \mathscr{F}_{t-1}) \times \mathbb{P}(\theta_j(t) \leq y_i, \forall j \neq 1 | \mathscr{F}_{t-1}, E_i^\theta(t)) \\ &= (1 - p_{i,t}) \times \mathbb{P}(\theta_j(t) \leq y_i, \forall j \neq 1 | \mathscr{F}_{t-1}, E_i^\theta(t)). \end{aligned} \tag{6}$$

The first inequality above follows from the fact that the event $\{i(t) = i | E_i^\theta(t)\}$ is a subset of the event $\{\theta_j(t), \forall j \leq y_i | E_i^\theta(t)\}$. Also, the first equality follows from the fact that the beta priors for different arms at any time t are independent of each other given the filtration $\mathscr{F}_{t-1}$. Conditioning on the event $E_i^\theta(t)$ retains the independence between $\theta_1(t)$ and $\theta_j(t), \forall j \neq 1$. Similarly, we have:

$$\begin{aligned} \mathbb{P}(i(t) = 1 | \mathscr{F}_{t-1}, E_i^\theta(t)) &\geq \mathbb{P}(\theta_1(t) > y_i \geq \theta_j(t), \forall j \neq 1 | \mathscr{F}_{t-1}, E_i^\theta(t)) \\ &= \mathbb{P}(\theta_1(t) > y_i | \mathscr{F}_{t-1}) \times \mathbb{P}(\theta_j(t) \leq y_i, \forall j \neq 1 | \mathscr{F}_{t-1}, E_i^\theta(t)) \\ &= p_{i,t} \times \mathbb{P}(\theta_j(t) \leq y_i, \forall j \neq 1 | \mathscr{F}_{t-1}, E_i^\theta(t)). \end{aligned} \tag{7}$$

Combining the inequalities (6) and (7) with (5), we get the lemma. □

Recall that τ_i denotes the time when the optimal arm 1 is transmitted the i^{th} time and $\tau_0 = 0$. Combining Lemma 1 with the first term on the RHS of (4):

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}(i(t) = i, E_i^{\theta}(t), E_i^{\mu}(t)) &= \sum_{t=1}^{T} \mathbb{E}[\mathbb{P}(i(t) = i, E_i^{\theta}(t), E_i^{\mu}(t) | \mathscr{F}_{t-1})] \\
&\leq \sum_{t=1}^{T} \mathbb{E}[\frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}(i(t) = 1, E_i^{\mu}(t), E_i^{\theta}(t) | \mathscr{F}_{t-1})] \\
&= \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[\frac{(1-p_{i,t})}{p_{i,t}} \mathbb{I}(i(t) = 1, E_i^{\mu}(t), E_i^{\theta}(t) | \mathscr{F}_{t-1})]] \\
&= \sum_{t=1}^{T} \mathbb{E}[\frac{(1-p_{i,t})}{p_{i,t}} \mathbb{I}(i(t) = 1, E_i^{\mu}(t), E_i^{\theta}(t))] \\
&\leq \sum_{k=0}^{T-1} \mathbb{E}[\frac{(1-p_{i,\tau_k+1})}{p_{i,\tau_k+1}} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}(i(t) = 1)] \\
&= \sum_{k=0}^{T-1} \mathbb{E}[\frac{1}{p_{i,\tau_k+1}} - 1]
\end{aligned} \tag{8}$$

where the first equality follows from the fact that given $\mathscr{F}_{t-1}$, $p_{i,t}$ is deterministic. The last inequality follows from the fact that $p_{i,t}$ only changes when the optimal arm 1 is played, hence it remains constant between time slots $\tau_k + 1$ and τ_{k+1}. Now, we bound the term $\mathbb{E}[\frac{1}{p_{i,\tau_k+1}}]$.

Lemma 2. *We have:*

$$\mathbb{E}[\frac{1}{p_{i,\tau_k+1}}] \leq \begin{cases} 1 + \frac{3}{\Delta_i'}, & \text{for } k < \frac{8}{\Delta_i'}, \\ 1 + \Theta(e^{-\Delta_i'^2 k/2} + \frac{1}{(k+1)\Delta_i'^2} e^{-D_i k} + \frac{1}{e^{\Delta_i'^2 k/4} - 1}), & \text{for } k \geq \frac{8}{\Delta_i'}, \end{cases}$$

where $\Delta_i' = \mu_1 - y_i, D_i = y_i \ln(\frac{y_i}{\mu_1}) + (1 - y_i) \ln(\frac{1-y_i}{1-\mu_1})$.

Proof. The proof of this lemma uses tight numerical estimates for partial Binomial sums. We don't present the proof here as it doesn't provide additional insights into the bigger picture of analyzing Algorithm 1. Interested reader can refer to Appendix B.3 in Agrawal and Goyal [2013] for details of the proof. □

Lemma 2 along with Equation (8) allows us to bound the first term in Equation (4). For the remaining two terms in Equation (4), we have the following lemmas:

Lemma 3.

$$\sum_{t=1}^{T} \mathbb{P}(i(t) = i, \overline{E_i^{\mu}(t)}) \leq \frac{1}{d(x_i, \mu_i)} + 1.$$

Proof. The proof for this lemma is a straightforward application of the Chernoff-Hoeffding bounds to $\hat{\mu}_i(t)$. We have:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}(i(t)=i, \overline{E_i^{\mu}(t)}) &\leq \mathbb{E}[\sum_{k=1}^{T} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}(i(t)=i)\mathbb{I}(\overline{E_i^{\mu}(t)})] \\
&\leq \mathbb{E}[\sum_{k=0}^{T-1} \mathbb{I}(\overline{E_i^{\mu}(\tau_k+1)}) \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}(i(t)=i)] \\
&\leq \mathbb{E}[\sum_{k=0}^{T-1} \mathbb{I}(\overline{E_i^{\mu}(\tau_k+1)})] \\
&\leq 1 + \mathbb{E}[\sum_{k=1}^{T-1} \mathbb{I}(\overline{E_i^{\mu}(\tau_k+1)})] \\
&\leq 1 + \exp(-kd(x_i,\mu_i)) \\
&\leq 1 + \frac{1}{d(x_i,\mu_i)},
\end{aligned}$$

where the second last inequality follows from the fact that the event $\overline{E_i^{\mu}(\tau_k+1)}$ is defined as the event that $\hat{\mu}_i(\tau_k+1) > x_i$. Using Chernoff-Hoeffding's bounds we get $\mathbb{P}(\hat{\mu}_i(\tau_k+1) > x_i) \leq \exp(-kd(x_i,\mu_i))$. □

Lemma 4.

$$\sum_{t=1}^{T} \mathbb{P}(i(t)=i, \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)) \leq L_i(T) + 1.$$

where $L_i(T) = \frac{\ln T}{d(x_i,y_i)}$.

Proof. The proof of this lemma follows from Chernoff-Hoeffding bounds and observing that $\theta_i(t)$ is well-concentrated around it's mean (i.e., $\overline{E_i^{\theta}(t)}$ is a low probability event) when the arm i has been played $L_i(t) = \frac{\ln T}{d(x_i,y_i)}$ times. Interested reader can refer to Appendix B.2 in Agrawal and Goyal [2013] for the details. □

Proof for Theorem 1: Now that we have the required lemmas to bound the three terms in Equation (4), we need to choose x_i, y_i appropriately in order to obtain the problem-dependent upper bound of Theorem 1. For any $0 < \epsilon \leq 1$, we set $x_i \in (\mu_i, \mu_1)$ such that $d(x_i,\mu_1) = \frac{d(\mu_i,\mu_1)}{1+\epsilon}$. Also, we set $y_i \in (x_i, \mu_1)$ such that $d(x_i,y_i) = \frac{d(x_i,\mu_1)}{1+\epsilon} = \frac{d(\mu_i,\mu_1)}{(1+\epsilon)^2}$. Therefore:

$$L_i(T) = \frac{\ln T}{d(x_i,y_i)} = (1+\epsilon)^2 \frac{\ln T}{d(\mu_i,\mu_1)}.$$

Also, by manipulating $d(x_i,\mu_1) = \frac{d(\mu_i,\mu_1)}{1+\epsilon}$, we can get:

$$x_i - \mu_i \geq \frac{\epsilon}{1+\epsilon} . \frac{d(\mu_i,\mu_1)}{\ln(\frac{\mu_1(1-\mu_i)}{\mu_i(1-\mu_1)})},$$

which implies $\frac{1}{d(x_i,\mu_i)} \leq \frac{1}{2(x_i-\mu_i)^2} = O(\frac{1}{\epsilon^2})$. Here big-Oh notation hides dependence on μ_is and Δ_is, assuming them as problem-dependent constants. Using the above and combining with Lemmas

1 - 4, and Equations (4) and (8):

$$\begin{aligned}\mathbb{E}[N_i(T+1)] &\leq \frac{24}{\Delta_i'^2} + \sum_{j=0}^{T-1} \Theta(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4} - 1}) \\ &\quad + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \\ &\leq \frac{24}{\Delta_i'^2} + \Theta(\frac{1}{\Delta_i'^2} + \frac{1}{\Delta_i'^2 D} + \frac{1}{\Delta_i'^4} + \frac{1}{\Delta_i'^2}) \\ &\quad + (1+\epsilon^2)\frac{\ln T}{d(\mu_i, \mu_1)} + O(\frac{1}{\epsilon^2}) \\ &= O(1) + (1+\epsilon^2)\frac{\ln T}{d(\mu_i, \mu_1)} + O(\frac{1}{\epsilon^2}).\end{aligned} \tag{9}$$

Again, the big-Oh notation hides the problem-dependent parameters μ_is and Δ_is as constants. Now, we can use the above inequality to get the final expected regret upper bound obtained in Theorem 1:

$$\begin{aligned}\mathbb{E}[R(T)] &= \sum_i \Delta_i \mathbb{E}[N_i(T+1)] \\ &\leq \sum_i (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)} \Delta_i + O(\frac{K}{\epsilon^2}) \\ &\leq \sum_i (1+\epsilon') \frac{\ln T}{d(\mu_i, \mu_1)} \Delta_i + O(\frac{K}{\epsilon'^2})\end{aligned}$$

where $\epsilon' = 3\epsilon$. This completes the proof for Theorem 1.

Proof for Theorem 2: The proof of Theorem 2 diverges from the proof of Theorem 1 only in the choice of x_i and y_i. In this case, we choose $x_i = \mu_i + \frac{\Delta_i}{3}$ and $y_i = \mu_1 - \frac{\Delta_i}{3}$. Therefore, we have: $\Delta_i'^2 = (\mu_1 - y_i)^2 = \frac{\Delta_i^2}{9}$. Using Pinsker's inequality, we get: $d(x_i, \mu_i) \geq 2(x_i - \mu_i)^2 = \frac{2\Delta_i^2}{9}$, and $d(x_i, y_i) \geq 2(x_i - y_i)^2 \geq \frac{2\Delta_i^2}{9}$. This gives us: $L_i(T) = \frac{\ln T}{d(x_i, y_i)} \leq \frac{9 \ln T}{2\Delta_i^2}$, and $\frac{1}{d(x_i, \mu_i)} \leq \frac{9}{2\Delta_i^2}$. Therefore, similar to the Equation (10) in the analysis of Theorem 1, we get:

$$\begin{aligned}\mathbb{E}[N_i(T+1)] &\leq \frac{24}{\Delta_i'^2} + \sum_{j=0}^{T-1} \Theta(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4} - 1}) \\ &\quad + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \\ &\leq \sum_{j=0}^{T-1} \Theta(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} + \frac{4}{\Delta_i'^2 j}) + O(\frac{\ln T}{\Delta_i^2}) \\ &= \Theta(\frac{1}{\Delta_i'^2} + \frac{\ln T}{\Delta_i'^2}) + O(\frac{\ln T}{\Delta_i^2}) = O(\frac{\ln T}{\Delta_i^2}).\end{aligned} \tag{10}$$

The remaining analysis is straightforward. From the above equation, for an arm i with $\Delta_i \geq \sqrt{\frac{K \ln T}{T}}$, expected regret $\Delta_i \mathbb{E}[N_i(T+1)] = O(\sqrt{\frac{T \ln T}{K}})$. For an arm i with $\Delta_i < \sqrt{\frac{K \ln T}{T}}$, the expected regret $\Delta_i \mathbb{E}[N_i(T+1)]$ is trivially $O(\sqrt{KT \ln T})$. Thus, the result in Theorem 2 follows.

3 Complex Online Problems

The standard MAB problem we focused on in this report is very useful in several practical scenarios. But, it is limited in its ability to model more complex reward and feedback structures. For instance, consider the following scenarios:

Scenario 1: The true expected reward vector $\bar{\mu}$ belongs to a subset $\psi \subset [0,1]^K$. If ψ is known to the player, can Thompson sampling algorithm exploit it?

Scenario 2: There is a complex relationship between the rewards observed and the arm chosen (instead of simply observing the reward of the arm played). Can Thompson sampling handle the complex reward feedback structure efficiently?

The answer to both of the above questions is **Yes**! Thompson sampling is a very powerful algorithm and naturally takes care of a lot of things. Recall the general structure behind the Thompson sampling algorithm:

1. At any time t, start with a prior distribution $\mathbb{P}_t(\theta)$ to sample a mean reward vector $\theta = (\theta_1, \theta_2, ..., \theta_K)$ from the set ψ.
2. Sample $\theta(t) \sim \mathbb{P}_t$.
3. Play the optimal arm for $\theta(t)$, i.e., play $i(t) = \arg\max_i \theta_i(t)$.
4. Receive the reward $r_{i(t)}(t)$.
5. Posterior update: $\mathbb{P}_{t+1}(\theta) \propto \mathbb{P}(r_{i(t)}(t)|\theta)\mathbb{P}_t(\theta)$.

Observe that the above structure allows for a non-trivial parameter set ψ. Also, almost any complex reward feedback can be captured in the posterior update step. Therefore, Thompson sampling can generally be used to handle more complex online sequential decision making problems. But, there are a few caveats:

1. The posterior update step is highly non-trivial for more complicated problems.
2. It is difficult to find a prior distribution that captures the structure in the parameter set, while being easy to update.

Although the above two challenges can be sometimes circumvented, it is still difficult to analyze and provide guarantees on the performance of Thompson sampling for complex online problems. The analysis which we presented in this report depends heavily on the fact that the problem we considered is the standard MAB problem. Interestingly, Gopalan et al. [2014] provide the first guarantees in their paper on the application of Thompson sampling to more general online decision making problems. They provide a theoretical guarantee based on the solution to a linear program that determines the worst-case regret the algorithm can have, computed across all possible trajectories it can take to eliminate the sub-optimal actions.

The major drawback of the analysis in Gopalan et al. [2014] is that it makes discretizing assumptions on the prior distribution and the problem parameters in order to make the analysis tractable. We omit their analysis from this report owing to space constraints, but recommend interested readers to refer to it. In practice, one cannot always discretize the prior or make assumptions on the granularity of the problem. Therefore, an open problem of interest is to prove theoretical guarantees on the performance of Thompson sampling algorithm for complex online problems, without restrictions on the priors it can use and without making assumptions on the underlying problem space.

4 Conclusion

In this report, we consider the standard multi-armed bandit problem and present the analysis of the Thompson sampling algorithm. We present optimal problem-dependent and near-optimal problem-independent upper bounds on its expected regret. We also presented a short discussion on Thompson sampling for more complex online sequential decision making problems and provide a challenging open problem to the reader.

References

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, pages 100–108, 2014.