Analysis of Thompson Sampling for the multi-armed bandit problem

Shipra Agrawal
Microsoft Research India
shipra@microsoft.com

Navin Goyal
Microsoft Research India
navingo@microsoft.com

Abstract

The multi-armed bandit problem is a popular model for studying exploration/exploitation trade-off in sequential decision problems. Many algorithms are now available for this well-studied problem. One of the earliest algorithms, given by W. R. Thompson, dates back to 1933. This algorithm, referred to as Thompson Sampling, is a natural Bayesian algorithm. The basic idea is to choose an arm to play according to its probability of being the best arm. Thompson Sampling algorithm has experimentally been shown to be close to optimal. In addition, it is efficient to implement and exhibits several desirable properties such as small regret for delayed feedback. However, theoretical understanding of this algorithm was quite limited. In this paper, for the first time, we show that Thompson Sampling algorithm achieves logarithmic expected regret for the stochastic multi-armed bandit problem. More precisely, for the two-armed stochastic bandit problem, the expected regret in time T is $O(\frac{\ln T}{\Delta} + \frac{1}{\Delta^3})$. And, for the N-armed stochastic bandit problem, the expected regret in time T is $O(\left[(\sum_{i=2}^N \frac{1}{\Delta_i^2})^2\right] \ln T)$. Our bounds are optimal but for the dependence on Δ_i and the constant factors in big-Oh.

1 Introduction

Multi-armed bandit (MAB) problem models the exploration/exploitation trade-off inherent in sequential decision problems. One of the early motivations for studying MAB problem was clinical trials: suppose that we have N different treatments of unknown efficacy for a certain disease. Patients arrive sequentially, and we must decide on a treatment to administer for each arriving patient. To make this decision, we could learn from how various treatments fared for the previous patients. After a sufficient number of trials, we may have a reasonable idea of which treatment is most effective, and from then on, we could administer that treatment for all the patients. However, initially, when there is no or very little information available, we need to *explore* and try each treatment sufficient number of times. We wish to do this exploration in such a way that we can find the best treatment and start *exploiting* it as soon as possible. The MAB problem is to decide how to choose the treatment for the next patient, given the outcomes of the treatments so far. Today, multi-armed bandit problem has a diverse set of applications some of which will be mentioned shortly.

Many versions and generalizations of the multi-armed bandit problem have been studied in the literature; in this paper we will consider a basic and well-studied version of this problem, referred to as stochastic multi-armed bandit problem. Among many algorithms available for the stochastic bandit problem, some popular ones include Upper Confidence Bound (UCB) family of algorithms (e.g., [7, 1]), which have good theoretical guarantees; and the algorithm by Gittins [3], which gives optimal strategy under known priors and geometric time-discounted rewards. In one of the earliest works on stochastic bandit problems, Thompson [11] proposed a natural randomized Bayesian algorithm to minimize regret. The basic idea is to play an arm with its probability of being the best arm. This algorithm is known as *Thompson Sampling* (TS), and it is a member of the family of *randomized probability matching* algorithms.

Recently, TS has attracted considerable attention. Several studies [5, 10, 2, 9] have empirically demonstrated the efficacy of Thompson Sampling: Scott [10] provides a detailed discussion of probability matching techniques in many general settings along with favorable empirical comparisons with other techniques. Chapelle and Li [2] demonstrate that empirically TS achieves regret comparable to the lower bound of [7]; and in applications like display advertising and news article recommendation, it is competitive to or better than popular methods such as UCB. In their experiments, TS is also more robust to delayed or batched feedback (in the above clinical trial example, delayed feedback would mean that the result of a treatment may become available after some time delay, but we are required to make immediate decisions for patients arriving in the mean time) than the other methods. A possible explanation may be that TS is a randomized algorithm and so it is unlikely to get trapped in an early bad decision during the delay. Microsoft's adPredictor [4] for CTR prediction of search ads on Bing also uses the idea of Thompson Sampling.

It has been suggested [2] that despite being easy to implement and being competitive to the state of the art methods, the reason TS is not very popular in literature could be its lack of strong theoretical analysis. Existing theoretical analyses [5, 8] provide weak guarantees, namely, a bound of $o(T)$ on expected regret in time T. Also, these existing works provide only asymptotic bounds, and do not bound the finite-time expected regret. In this paper, for the first time, we provide a logarithmic bound on expected regret of TS algorithm in time T that is close to the lower bound of [7]. Further, we show that TS achieves logarithmic regret uniformly over time, rather than only asymptotically. Before stating our results, we describe the MAB problem and the TS algorithm formally.

1.1 The multi-armed bandit problem

We consider the stochastic multi-armed bandit (MAB) problem: We are given a slot machine with N arms; at each time step $t = 1, 2, 3, \ldots$, one of the N arms must be chosen to be played. Each arm i, when played, yields a random real-valued reward according to some fixed (unknown) distribution with support in $[0, 1]$. The random reward obtained from playing an arm repeatedly are i.i.d. and independent of the plays of the other arms. The reward is observed immediately after playing the arm.

An algorithm for the MAB problem must decide which arm to play at each time step t, based on the outcomes of the previous $t-1$ plays. Let μ_i denote the (unknown) expected reward for arm i. A popular goal is to maximize the expected total reward in time T, i.e., $\mathbb{E}[\sum_{t=1}^{T} \mu_{i(t)}]$, where $i(t)$ is the arm played in step t, and the expectation is over the random choices of $i(t)$ made by the algorithm. It is more convenient to work with the equivalent measure of expected total *regret*: the amount we lose because of not playing optimal arm in each step. To formally define regret, let us introduce some notation. Let $\mu^* := \max_i \mu_i$, and $\Delta_i := \mu^* - \mu_i$. Also, let $k_i(t)$ denote the number of times arm i has been played up to step $t-1$. Then the expected total regret in time T is given by

$$\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}[\sum_{t=1}^{T}(\mu^* - \mu_{i(t)})] = \sum_i \Delta_i \cdot \mathbb{E}[k_i(T)].$$

Other performance measures include PAC-style guarantees; we do not consider those measures here.

1.2 Thompson Sampling

In the most general setting, Thompson Sampling can be described as a natural Bayesian algorithm that plays an arm according to its probability of being the best arm. For simplicity of discussion, we first provide the details of this algorithm for the Bernoulli bandit problem, i.e. when the rewards are either 0 or 1, and for arm i the probability of success (reward $=1$) is μ_i. This description of Thompson Sampling follows closely that of Chapelle and Li [2]. Next, we propose a simple new extension of this algorithm to general reward distributions with support $[0, 1]$, which will allow us to seamlessly extend our analysis for Bernoulli bandits to general stochastic bandit problem.

The algorithm for Bernoulli bandits maintains Bayesian priors on the Bernoulli means μ_i's. Beta distribution turns out to be a very convenient choice of priors for Bernoulli rewards. Let us briefly recall that beta distributions form a family of continuous probability distributions on the interval $(0, 1)$. The pdf of $Beta(\alpha, \beta)$, the beta distribution with parameters $\alpha > 0$, $\beta > 0$, is given by $f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$. The mean of $Beta(\alpha, \beta)$ is $\alpha/(\alpha+\beta)$; and as is apparent from the pdf, higher the α, β, tighter is the concentration of $Beta(\alpha, \beta)$ around the mean. Beta distribution is useful for Bernoulli rewards because if the prior is a beta distribution, then so is the posterior.

Initially, the algorithm assumes arm i to have prior $Beta(1, 1)$ on μ_i, which is natural because $Beta(1, 1)$ is the uniform distribution on $(0, 1)$. At time t, having observed $S_i(t)$ successes (reward $= 1$) and $F_i(t)$ failures (reward $= 0$) in $k_i(t) = S_i(t) + F_i(t)$ plays of arm i, the algorithm assumes a $Beta(S_i(t)+1, F_i(t)+1)$ prior on μ_i. Note that the beta distribution with parameters $S_i(t) + 1, F_i(t) + 1$ is the posterior distribution of μ_i after observing $S_i(t)$ successes (with probability μ_i of success) and $F_i(t)$ failures (with probability $1 - \mu_i$ of failure). The algorithm then samples from this posterior distribution of the μ_i's, and plays an arm according to the probability of its mean being the largest.

We summarize the Thompson Sampling algorithm below.

Algorithm 1 Thompson Sampling for Bernoulli bandits

$S_i = 0, F_i = 0$.
for $t = 1, 2, \ldots,$ **do**
  For each arm $i = 1, \ldots, N$, generate $\theta_i(t)$ from the beta distribution with parameters $(S_i + 1, F_i + 1)$.
  Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward r.
  if $r = 1$, then $S_i = S_i + 1$, else $F_i = F_i + 1$.
end for

For general stochastic bandits, the rewards for arm i are generated from a distribution with support $[0, 1]$ and mean μ_i. Let $\tilde{r}_t \in [0, 1]$ denote the reward observed at time t. We modify the Thompson Sampling algorithm to perform a Bernoulli trial with probability with success probability $\tilde{r}_t$ after observing the reward at time t. The variables $\{S_i(t), F_i(t)\}$ now denote the number of successes and failures in these Bernoulli trials. The remaining algorithm is essentially the same as for Bernoulli bandits. Below is a precise description of the generalized algorithm.

Algorithm 2 Thompson Sampling for general stochastic bandits

$S_i = 0, F_i = 0$.
for $t = 1, 2, \ldots,$ **do**
  For each arm $i = 1, \ldots, N$, generate $\theta_i(t)$ from the beta distribution with parameters $(S_i + 1, F_i + 1)$.
  Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward $\tilde{r}$.
  Perform a Bernoulli trial with success probability $\tilde{r}$ and observe output r.
  if $r = 1$, then $S_i = S_i + 1$, else $F_i = F_i + 1$.
end for

We observe that the probability of observing $r = 1$ on playing an arm i in the new generalized algorithm is equal to the mean reward μ_i. Let f_i denote the pdf of reward distribution for arm i. Then, on playing arm i,

$$\Pr(r = 1) = \int_0^1 \tilde{r} f_i(\tilde{r}) d\tilde{r} = \mu_i.$$

Thus, the probability of observing $r = 1$ is same, and $S_i(t), F_i(t)$ evolve exactly in the same way as in the case of Bernoulli bandits with mean μ_i. This allows us to replace, for analysis purpose, the general stochastic bandits with Bernoulli bandits with the same means. We use this observation to confine the proofs in this paper to the case of Bernoulli bandits only.

1.3 Our results

In this article, we bound the *finite time* expected regret of Thompson Sampling. From now on we will assume that the first arm is the unique optimal arm, i.e., $\mu^* = \mu_1 > \arg\max_{i \neq 1} \mu_i$. Assuming that the first arm is an optimal arm is a matter of convenience for stating the results and for the analysis. The assumption of *unique* optimal arm is also without loss of generality, since adding more arms with $\mu_i = \mu^*$ can only decrease the expected regret; details of this argument are provided in Appendix A.

Theorem 1. *For the two-armed stochastic bandit problem ($N = 2$), Thompson Sampling algorithm has expected regret*

$$\mathbb{E}[\mathcal{R}(T)] \le \frac{32}{\Delta}\ln T + \frac{48}{\Delta^3} + 18\Delta$$

in time T, where $\Delta = \mu_1 - \mu_2$.

Theorem 2. *For the N-armed stochastic bandit problem, Thompson Sampling algorithm has expected regret*

$$\mathbb{E}[\mathcal{R}(T)] \le \left[768(\sum_{i=2}^{N}\frac{1}{\Delta_i^2})^2 + 192(\sum_{i=2}^{N}\frac{1}{\Delta_i^2}) + 64(\sum_{i=2}^{N}\frac{1}{\Delta_i})\right]\ln T + 192\left(\sum_{i=2}^{N}\frac{1}{\Delta_i^2}\right) + 52(N-1)$$

in time T, where $\Delta_i = \mu_1 - \mu_i$.

We remark that we have not attempted to optimize constants in the above theorems in the interest of readability. Let us contrast our bounds with the previous work. Lai and Robbins [7] proved the following lower bound on regret of any bandit algorithm:

$$\mathbb{E}[\mathcal{R}(T)] \ge \left[\sum_{i=2}^{N}\frac{\Delta_i}{D(\mu_i||\mu)} + o(1)\right]\ln T,$$

where D denotes the KL divergence. They also gave algorithms asymptotically achieving this guarantee, though unfortunately their algorithms are not efficient. Auer et al. [1] gave the UCB1 algorithm, which is efficient and achieves the following bound:

$$\mathbb{E}[\mathcal{R}(T)] \le \left[8\sum_{i=2}^{N}\frac{1}{\Delta_i}\right]\ln T + (1+\pi^2/3)\left(\sum_{i=2}^{N}\Delta_i\right).$$

For many settings of the parameters, the bound of Auer et al. is not far from the lower bound of Lai and Robbins. Our bounds are optimal in terms of dependence on T, but inferior in terms of the constant factors and dependence on Δ. We note that for the two-armed case our bound closely matches the bound of Auer et al. For the N-armed setting, the exponent of Δ's in our bound is basically 4 compared to the exponent 1 for UCB1.

2 Proof Techniques

In this section, we give an informal description of the techniques involved in our analysis. We hope that this will aid in reading the proofs, though this section is not essential for the sequel. We assume that all arms are Bernoulli arms, and that the first arm is the unique optimal arm. As explained in the previous sections, these assumptions are without loss of generality.

Main technical difficulties. Thompson Sampling is a randomized algorithm which achieves exploration by choosing to play the arm with best sampled mean, among those generated from beta distributions around the respective empirical means. The beta distribution becomes more and more concentrated around the empirical mean as the number of plays of an arm increases. This randomized setting is unlike the algorithms in UCB family, which achieve exploration by adding a *deterministic, non-negative* bias inversely proportional to the number of plays, to the observed empirical means. Analysis of TS poses difficulties that seem to require new ideas.

For example, following general line of reasoning is used to analyze regret of UCB like algorithms in two-arms setting [1]: once the second arm has been played sufficient number of times, its empirical mean is tightly concentrated around its actual mean. If the first arm has been played sufficiently large number of times by then, it will have an empirical mean close to its actual mean and larger than that of the second arm. Otherwise, if it has been played small number of times, its non-negative bias term will be large. Consequently, once the second arm has been played sufficient number of times, it will be played with very small probability (inverse polynomial of time) *regardless of the number of times the first arm has been played so far.*

However, for Thompson Sampling, if the number of previous plays of the first arm is small, then the probability of playing the second arm could be as large as a constant even if it has already been played large number of times. For instance, if the first arm has not been played at all, then $\theta_1(t)$ is a uniform random variable, and thus $\theta_1(t) < \theta_2(t)$ with probability $\theta_2(t) \approx \mu_2$. As a result, in our analysis we need to carefully consider the distribution of the number of previous plays of the first arm, in order to bound the probability of playing the second arm.

The observation just mentioned also points to a challenge in extending the analysis of TS for two-armed bandit to the general N-armed bandit setting. One might consider analyzing the regret in the N-armed case by considering only two arms at a time—the first arm and one of the suboptimal arms. We could use the observation that the probability of playing a suboptimal arm is bounded by the probability of it exceeding the first arm. However, this probability also depends on the number of previous plays of the two arms, which in turn depend on the plays of the other arms. Again, Auer et al.[1], in their analysis of UCB algorithm, overcome this difficulty by bounding this probability for *all possible numbers of previous plays* of the first arm, and large enough plays of the suboptimal arm. For Thompson Sampling, due to the observation made earlier, the (distribution of the) number of previous plays of the first arm needs to be carefully accounted for, which in turn requires considering all the arms at the same time, thereby leading to a more involved analysis.

Proof outline for two arms setting. Let us first consider the special case of two arms which is simpler than the general N arms case. Firstly, we note that it is sufficient to bound the regret incurred during the time steps *after* the second arm has been played $L = 16(\ln T)/\Delta^2$ times. The expected regret before this event is bounded by $16(\ln T)/\Delta$ because only the plays of the second arm produce an expected regret of Δ; regret is 0 when the first arm is played. Next, we observe that after the second arm has been played L times, the following happens with high probability: the empirical average reward of the second arm from each play is very close to its actual expected reward μ_2, and its beta distribution is tightly concentrated around μ_2. This means that, thereafter, the first arm would be played at time t if $\theta_1(t)$ turns out to be greater than (roughly) μ_2. This observation allows us to model the number of steps between two consecutive plays of the first arm as a geometric random variable with parameter close to $\Pr[\theta_1(t) > \mu_2]$. To be more precise, given that there have been j plays of the first arm with $s(j)$ successes and $f(j) = j - s(j)$ failures, we want to estimate the expected number of steps before the first arm is played again (not including the steps in which the first arm is played). This is modeled by a geometric random variable $X(j, s(j), \mu_2)$ with parameter $\Pr[\theta_1 > \mu_2]$, where θ_1 has distribution $\text{Beta}(s(j)+1, j-s(j)+1)$, and thus $\mathbb{E}[X(j, s(j), \mu_2)|s(j)] = 1/\Pr[\theta_1 > \mu_2] - 1$. To bound the overall expected number of steps between the j^{th} and $(j+1)^{th}$ play of the first arm, we need to take into account the distribution of the number of successes $s(j)$. For large j, we use Chernoff–Hoeffding bounds to say that $s(j)/j \approx \mu_1$ with high probability, and moreover θ_1 is concentrated around its mean, and thus we get a good estimate of $\mathbb{E}[\mathbb{E}[X(j, s(j), \mu_2)|s(j)]]$. However, for small j we do not have such concentration, and it requires a delicate computation to get a bound on $\mathbb{E}[\mathbb{E}[X(j, s(j), \mu_2)|s(j)]]$. The resulting bound on the expected number of steps between consecutive plays of the first arm bounds the expected number of plays of the second arm, to yield a good bound on the regret for the two-arms setting.

Proof outline for N arms setting. At any step t, we divide the set of suboptimal arms into two subsets: *saturated* and *unsaturated*. The set $C(t)$ of saturated arms at time t consists of arms a that have already been played a sufficient number ($L_a = 16(\ln T)/\Delta_a^2$) of times, so that with high probability, $\theta_a(t)$ is tightly concentrated around μ_a. As earlier, we try to estimate the number of steps between two consecutive plays of the first arm. After j^{th} play, the $(j+1)^{th}$ play of first arm will occur at the earliest time t such that $\theta_1(t) > \theta_i(t), \forall i \neq 1$. The number of steps before $\theta_1(t)$ is greater than $\theta_a(t)$ of all saturated arms $a \in C(t)$ can be analyzed using a geometric random variable with parameter close to $\Pr(\theta_1 \geq \max_{a \in C(t)} \mu_a)$, as before. However, even if $\theta_1(t)$ is greater than the $\theta_a(t)$ of all saturated arms $a \in C(t)$, it may not get played due to play of an unsaturated arm u with a greater $\theta_u(t)$. Call this event an "interruption" by unsaturated arms. We show that if there have been j plays of first arm with $s(j)$ successes, the expected number of steps until the $(j+1)^{th}$ play can be upper bounded by the product of the expected value of a geometric random variable similar to $X(j, s(j), \max_a \mu_a)$ defined earlier, and the number of interruptions by the unsaturated arms. Now, the total number of interruptions by unsaturated arms is bounded by $\sum_{u=2}^{N} L_u$ (since an arm u becomes saturated after L_u plays). The actual number of interruptions is hard to analyze due to the high variability in the parameters of the unsaturated arms. We derive our bound assuming the worst case allocation of these $\sum_u L_u$ interruptions. This step in the analysis is the main source of the high exponent of Δ in our regret bound for the N-armed case compared to the two-armed case.

3 Regret bound for the two-armed bandit problem

In this section, we present a proof of Theorem 1, our result for the two-armed bandit problem. Recall our assumption that all arm have Bernoulli distribution on rewards, and that the first arm is the unique optimal arm.

Let random variable j_0 denote the number of plays of the first arm until $L = 16(\ln T)/\Delta^2$ plays of the second arm. Also, let random variable Y_j measure the number of time steps between the j^{th} and $(j+1)^{th}$ plays of the first arm (not counting the steps in which the j^{th} and $(j+1)^{th}$) plays happened), and let $s(j)$ denote the number of successes in the first j plays of the first arm. Then the expected number of plays of the second arm in time T is bounded by

$$\mathbb{E}[k_2(T)] \leq L + \mathbb{E}[\textstyle\sum_{j=j_0}^{T-1} Y_j].$$

To understand the expectation of Y_j, it will be useful to define another random variable $X(j, s, y)$ as follows. We perform the following experiment until it succeeds: check if a Beta$(s+1, j-s+1)$ distributed random variable exceeds a threshold y. For each experiment, we generate the beta-distributed r.v. independently of the previous ones. Now define $X(j, s, y)$ to be the number of trials *before* the experiment succeeds. Thus, $X(j, s, y)$ takes non-negative integer values, and is a geometric random variable with parameter (success probability) $1 - F^{beta}_{s+1,j-s+1}(y)$. Here $F^{beta}_{\alpha,\beta}$ denotes the cdf of the beta distribution with parameters α, β.

We will relate Y and X shortly. The following lemma provides a handle on the expectation of X.

Lemma 1. *For all non-negative integers $j, s \leq j$, and for all $y \in [0, 1]$,*

$$\mathbb{E}[X(j, s, y)] = \frac{1}{F^B_{j+1,y}(s)} - 1,$$

where $F^B_{n,p}$ denotes the cdf of the binomial distribution with parameters (n, p).

Proof. By the well-known formula for the expectation of a geometric random variable and the definition of X we have, $\mathbb{E}[X(j, s, y)] = \frac{1}{1-F^{beta}_{s+1,j-s+1}(y)} - 1$ (The additive -1 is there because we do not count the final step where the Beta r.v. is greater than y.) The lemma then follows from Fact 1 in Appendix B. □

In order to bound expected value of Y using expected value of X, we will condition on an event $E_2(T)$, which we define to hold iff $\theta_2(t) \leq \mu_2 + \frac{\Delta}{2}$ for all time $t \in [1, T]$ such that the second arm has already been played at least L times before time t. The following lemma can be proven using Fact 1 and the Chernoff–Hoeffding bounds.

Lemma 2. $\Pr(E_2(T)) \geq 1 - \frac{2}{T}$.

Proof. Refer to Appendix C.1. □

Recall that Y_j was defined as the number of steps before $\theta_1(t) > \theta_2(t)$ for the first time after the j^{th} play of the first arm. Now, under event $E_2(T)$, $\theta_2(t) \leq \mu_2 + \Delta/2$, for all time t after L plays of the second arm. Therefore, given that event $E_2(T)$ holds, and that the number of successes in j trials of first arm is $s(j)$, Y_j for $j \geq j_0$ is stochastically dominated by geometric random variable $X(j, s(j), \mu_2 + \frac{\Delta}{2})$. Also, it is bounded by T. If the event $E_2(T)$ does not hold, we can bound $\sum_j Y_j$ by the trivial upper bound of T. Using these observations,

$$\begin{aligned}
\mathbb{E}[\sum_{j=j_0}^{T-1} Y_j] &\leq \mathbb{E}[\sum_{j=j_0}^{T-1} \mathbb{E}[\ Y_j | j_0, s(j), E_2(T)\ ]\ ] + \overline{\Pr(E_2(T))} \cdot T \\
&\leq \mathbb{E}[\sum_{j=j_0}^{T-1} \min\{\mathbb{E}[X(j, s(j), \mu_2 + \frac{\Delta}{2}) \mid s(j)], T\}] + \frac{2}{T} \cdot T \\
&\leq \mathbb{E}[\sum_{j=0}^{T-1} \min\{\mathbb{E}[X(j, s(j), \mu_2 + \frac{\Delta}{2}) \mid s(j)], T\}] + 2.
\end{aligned}$$

Lemma 3. *Consider any positive $y < \mu_1$, and let $\Delta' = \mu_1 - y$. Also, let $R = \frac{\mu_1(1-y)}{y(1-\mu_1)} > 1$, and let D denote the KL-divergence between μ_1 and y, i.e. $D = y \ln \frac{y}{\mu_1} + (1-y)\ln\frac{1-y}{1-\mu_1}$. Then,*

$$\mathbb{E}\left[\min\{\ \mathbb{E}[X(j, s(j), y) \mid s(j)],\ T\}\right] \quad \leq \quad \begin{cases} 1 + \dfrac{2}{1-y} + \dfrac{\mu_1}{\Delta'} e^{-Dj} & j < \frac{y}{D}\ln R, \\ 1 + \dfrac{R^y}{1-y} e^{-Dj} + \dfrac{\mu_1}{\Delta'} e^{-Dj} & \frac{y}{D}\ln R \leq j < \frac{4 \ln T}{\Delta'^2}, \\ \dfrac{16}{T} & j \geq \frac{4\ln T}{\Delta'^2}, \end{cases}$$

where the outer expectation is taken over $s(j)$ distributed as Binomial(j, μ_1).

Proof. Using Lemma 1, the expected value of $X(j, s(j), y)$ for any given $s(j)$,

$$\mathbb{E}[X(j, s(j), y) \mid s(j)] = \frac{1}{F^B_{j+1,y}(s(j))} - 1.$$

Case of large j: First, we consider the case of large j, i.e. when $j \geq 4(\ln T)/\Delta'^2$. Then, by simple application of Chernoff–Hoeffding bounds (refer to Fact 3 and Lemma 5), we can derive that for any $s \geq (y + \frac{\Delta'}{2})j$,

$$F^B_{j+1,y}(s) \geq F^B_{j+1,y}(yj + \frac{\Delta' j}{2}) \geq 1 - \frac{e^{4\Delta'/2}}{e^{2j\Delta'^2/4}} \geq 1 - \frac{e^{2\Delta'}}{T^2} \geq 1 - \frac{8}{T^2},$$

giving that for $s \geq y(j + \frac{\Delta'}{2})$, $\mathbb{E}[X(j+1, s, y)] \leq \frac{1}{(1-\frac{8}{T^2})} - 1$.

Again using Chernoff–Hoeffding bounds, the probability that $s(j)$ takes values smaller than $(y + \frac{\Delta'}{2})j$ can be bounded as,

$$F^B_{j,\mu_1}(yj + \frac{\Delta' j}{2}) = F^B_{j,\mu_1}(\mu_1 j - \frac{\Delta' j}{2}) \leq e^{-2j\frac{\Delta'^2}{4}} \leq \frac{1}{T^2} < \frac{8}{T^2}.$$

For these values of $s(j)$, we will use the upper bound of T. Thus,

$$\mathbb{E}[\min\{\mathbb{E}[X(j, s(j), y)|s(j)], T\}] \quad \leq \quad (1 - 8/T^2) \cdot \left(\frac{1}{(1 - 8/T^2)} - 1\right) + \frac{8}{T^2} \cdot T \leq \frac{16}{T}.$$

Case of small j: For small j, the argument is more delicate. We use,

$$\mathbb{E}[\mathbb{E}[X(j, s(j), y)|s(j)]] = \mathbb{E}[\frac{1}{F^B_{j+1,y}(s(j))} - 1] = \sum_{s=0}^{j} \frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} - 1, \tag{1}$$

where f^B_{j,μ_1} denotes pdf of the Binomial(j, μ_1) distribution. We use the observation that for $s \geq \lceil y(j+1) \rceil$, $F^B_{j+1,y}(s) \geq 1/2$. This is because the median of a Binomial(n, p) distribution is either $\lfloor np \rfloor$ or $\lceil np \rceil$ [6]. Therefore,

$$\sum_{s=\lceil y(j+1)\rceil}^{j} \frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} \leq 2. \tag{2}$$

For small s, i.e., $s \le \lfloor yj \rfloor$, we use $F^B_{j+1,y}(s) = (1-y)F^B_{j,y}(s) + yF_{j,y}(s-1) \ge (1-y)F^B_{j,y}(s)$ and $F^B_{j,y}(s) \ge f^B_{j,y}(s)$, to get

$$
\begin{aligned}
\sum_{s=0}^{\lfloor yj \rfloor} \frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} &\le \sum_{s=0}^{\lfloor yj \rfloor} \frac{1}{(1-y)} \frac{f^B_{j,\mu_1}(s)}{f^B_{j,y}(s)} \\
&= \sum_{s=0}^{\lfloor yj \rfloor} \frac{1}{(1-y)} \frac{\mu_1^s (1-\mu_1)^{j-s}}{y^s (1-y)^{j-s}} \\
&= \sum_{s=0}^{\lfloor yj \rfloor} \frac{1}{(1-y)} R^s \frac{(1-\mu_1)^j}{(1-y)^j} \\
&= \frac{1}{(1-y)} \left(\frac{R^{\lfloor yj \rfloor + 1} - 1}{R-1} \right) \frac{(1-\mu_1)^j}{(1-y)^j} \\
&\le \frac{1}{(1-y)} \frac{R}{R-1} \frac{\mu_1^{yj} (1-\mu_1)^{(j-yj)}}{y^{yj}(1-y)^{j-yj}} \\
&= \frac{\mu_1}{\mu_1 - y} e^{-Dj} = \frac{\mu_1}{\Delta'} e^{-Dj}. \qquad (3)
\end{aligned}
$$

If $\lfloor yj \rfloor < \lceil yj \rceil < \lceil y(j+1) \rceil$, then we need to additionally consider $s = \lceil yj \rceil$. Note, however, that in this case $\lceil yj \rceil \le yj + y$. For $s = \lceil yj \rceil$,

$$
\begin{aligned}
\frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} &\le \frac{1}{(1-y)F^B_{j,y}(s)} \\
&\le \frac{2}{1-y}. \qquad (4)
\end{aligned}
$$

Alternatively, we can use the following bound for $s = \lceil yj \rceil$,

$$
\begin{aligned}
\frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} &\le \frac{1}{(1-y)} \frac{f^B_{j,\mu_1}(s)}{F^B_{j,y}(s)} \\
&\le \frac{1}{(1-y)} \frac{f^B_{j,\mu_1}(s)}{f^B_{j,y}(s)} \\
&\le \frac{1}{(1-y)} R^s \left(\frac{1-\mu_1}{1-y} \right)^j \\
&\le \frac{1}{(1-y)} R^{yj+y} \left(\frac{1-\mu_1}{1-y} \right)^j \qquad \text{(because } s = \lceil yj \rceil \le yj + y\text{)} \\
&\le \frac{R^y}{(1-y)} e^{-Dj}. \qquad (5)
\end{aligned}
$$

Next, we substitute the bounds from (2)-(5) in Equation (1) to get the result in the lemma. In this substitution, for $s = \lceil yj \rceil$, we use the bound in Equation (4) when $j < \frac{y}{D} \ln R$, and the bound in Equation (5) when $j \ge \frac{y}{D} \ln R$.

□

Using Lemma 3 for $y = \mu_2 + \Delta/2$, and $\Delta' = \Delta/2$, we can bound the expected number of plays of the second arm as:

$$
\begin{aligned}
\mathbb{E}[k_2(T)] &= L + \mathbb{E}[\sum_{j=j_0}^{T-1} Y_j] \\
&\leq L + \sum_{j=0}^{T-1} \mathbb{E}[\min\{\mathbb{E}[X(j, s(j), \mu_2 + \frac{\Delta}{2}) \mid s(j)], \quad T\}\ ] + 2 \\
&\leq L + \frac{4 \ln T}{\Delta'^2} + \sum_{j=0}^{4(\ln T)/\Delta'^2 - 1} \frac{\mu_1}{\Delta'} e^{-Dj} + \left(\frac{y}{D} \ln R\right) \frac{2}{1-y} + \sum_{j=\frac{y}{D}\ln R}^{4(\ln T)/\Delta'^2 - 1} \frac{R^y e^{-Dj}}{1-y} + \frac{16}{T} \cdot T + 2 \\
&= L + \frac{4 \ln T}{\Delta'^2} + \sum_{j=0}^{4(\ln T)/\Delta'^2 - 1} \frac{\mu_1}{\Delta'} e^{-Dj} + \frac{y}{D} \ln R \cdot \frac{2}{(1-y)} + \sum_{j=0}^{4\ln\Delta'^2 - \frac{y}{D}\ln R - 1} \frac{1}{1-y} e^{-Dj} + 18 \\
&\leq L + \frac{4 \ln T}{\Delta'^2} + \frac{y}{D} \ln R \cdot \frac{2}{\Delta'} + \sum_{j=0}^{T-1} \frac{(\mu_1 + 1)}{\Delta'} e^{-Dj} + 18 \\
&\overset{(*)}{\leq} L + \frac{4 \ln T}{\Delta'^2} + \frac{D+1}{\Delta' D} \cdot \frac{2}{\Delta'} + \frac{2}{\Delta'} \frac{2}{(\min\{D, 1\})} + 18 \\
&\overset{(**)}{\leq} L + \frac{4 \ln T}{\Delta'^2} + \frac{2}{\Delta'^2} + \frac{1}{\Delta'^4} + \frac{4}{\Delta'^3} + 18 \\
&= L + \frac{16 \ln T}{\Delta^2} + \frac{8}{\Delta^2} + \frac{16}{\Delta^4} + \frac{32}{\Delta^3} + 18 \\
&\leq \frac{32 \ln T}{\Delta^2} + \frac{48}{\Delta^4} + 18.
\end{aligned}
$$

The step marked $(*)$ is obtained using following derivations.

$$y \ln R = y \ln \frac{\mu_1(1-y)}{y(1-\mu_1)} = y \ln \frac{\mu_1}{y} + y \ln \frac{(1-y)}{(1-\mu_1)} \leq \mu_1 + \frac{y}{1-y}(D - y \ln \frac{y}{\mu_1}) \leq 1 + \frac{y}{1-y}(D + \mu_1) \leq \frac{D+1}{\Delta'}.$$

And, since $D \geq 0$ (Gibbs' inequality),

$$\sum_{j \geq 0} e^{-Dj} = \frac{1}{1 - e^{-D}} \leq \max\{\frac{2}{D}, \frac{e}{e-1}\} \leq \frac{2}{\min\{D, 1\}}.$$

And, $(**)$ uses Pinsker's inequality to obtain $D \geq 2\Delta'^2$.

This gives a regret bound of

$$\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}[\Delta \cdot\ k_2(T)] \leq \left(\frac{32 \ln T}{\Delta} + \frac{48}{\Delta^3} + 18\Delta\right).$$

4 Regret bound for the N-armed bandit problem

In this section, we prove Theorem 2, our result for the N-armed bandit problem. Again, we assume that all arms have Bernoulli distribution on rewards, and that the first arm is the unique optimal arm.

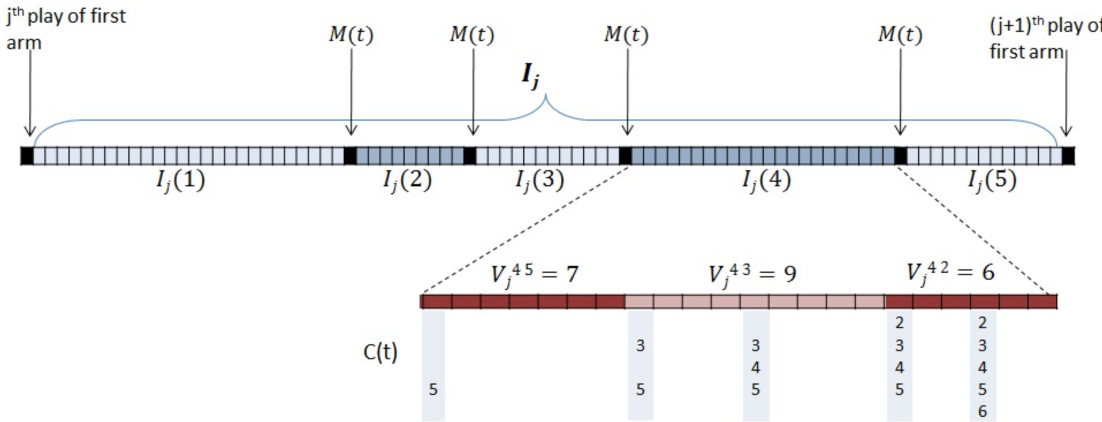

Figure 1: Interval I_j

At every time step t, we divide the set of suboptimal arms into saturated and unsaturated arms. We say that an arm $i \neq 1$ is in the saturated set $C(t)$ at time t, if it has been played at least $L_i := \frac{16 \ln T}{\Delta_i^2}$ times before time t. First, we will bound the regret due to playing saturated suboptimal arms. For this we bound the number of time steps between two consecutive plays of the first arm in which $\theta_i(t)$ of some saturated arm i exceeds $\theta_1(t)$.

In the following, when we talk about an interval of time, we mean a set of contiguous time steps. Let r.v. I_j denote the interval between (and excluding) the j^{th} and $(j+1)^{th}$ play of first arm. We say that event $M(t)$ holds at time t, if $\theta_1(t)$ exceeds $\theta_i(t)$ of all the saturated arms, i.e.,

$$M(t) : \theta_1(t) > \max_{i \in C(t)} \theta_i(t). \tag{6}$$

Let r.v. γ_j denote the number of occurrences of event $M(t)$ in interval I_j.

$$\gamma_j = |\{t \in I_j : M(t) = 1\}|. \tag{7}$$

Let r.v. $I_j(\ell)$ denote the sub-interval of I_j between the $(\ell-1)^{th}$ and ℓ^{th} occurrence of event $M(t)$ in I_j (excluding the time steps at which the event $M(t)$ occurs). $I_j(1)$ denotes the sub-interval before the first occurrence of event $M(t)$ in I_j; and $I_j(\gamma_j + 1)$ denotes the sub-interval after the last occurrence of event $M(t)$ in I_j.

Figure 1 shows an example of interval I_j along with sub-intervals $I_j(\ell)$; in this figure $\gamma_j = 4$.

Observe that since a saturated arm i can be played at step t only if $\theta_i(t)$ is greater than $\theta_1(t)$, saturated arms can only be played in the time steps belonging to intervals $I_j(\ell), \ell = 1, \ldots, \gamma_j + 1$. (Notice, however, that it is possible that at a step $t \in I_j(\ell)$ no saturated arm is played because an unsaturated arm has the greatest $\theta(t)$.) Therefore, the number of plays of saturated arms in interval I_j is at most

$$\sum_{\ell=1}^{\gamma_j+1} |I_j(\ell)|.$$

Now, let $V_j^{\ell,a}$ denote the number of steps in $I_j(\ell)$, for which a is the best saturated arm, i.e.

$$V_j^{\ell,a} = |\{t \in I_j(\ell) : \mu_a = \max_{i \in C(t)} \mu_i\}|. \tag{8}$$

(resolve the ties for best saturated arm using an arbitrary, but fixed, ordering on arms). Note that $|I_j(\ell)| = \sum_{a=2}^{N} V_j^{\ell,a}$. In Figure 1, we illustrate this notation by showing steps $\{V_j^{4,a}\}$ for interval $I_j(4)$. In the example shown, we assume that $\mu_1 > \mu_2 > \cdots > \mu_6$, and that the suboptimal arms got added to the saturated set $C(t)$ in order $5, 3, 4, 2, 6$, so that initially 5 is the best saturated arm, then 3 is the best saturated arm, and finally 2 is the best saturated arm.

Next, we will show that, with high probability, the regret due to playing a saturated arm during one of the $V_j^{\ell,a}$ steps is at most $3\Delta_a$. The idea is that since saturated arms have their $\theta_i(t)$s tightly concentrated around their means μ_i, so with high probability, either the arm with the highest mean, i.e., the best saturated arm a, or an arm with mean very close to μ_a will be chosen to be played during these $V_j^{\ell,a}$ steps.

More precisely, define $E(T)$ to be the event that for all $t \in [1, T]$, and for all $i \in C(t)$, $\mu_i - \Delta_i/2 \leq \theta_i(t) \leq \mu_i + \Delta_i/2$.

Lemma 4. $\Pr(E(T)) \geq 1 - \frac{4(N-1)}{T}$.

Proof. Refer to Appendix C.2. □

Given that $E(T)$ holds, if a saturated arm i is played at a time t among one of the $V_j^{\ell,a}$ steps, then,

$$\mu_i + \Delta_i/2 \geq \theta_i(t) \geq \theta_a(t) \geq \mu_a - \Delta_a/2,$$

which implies that

$$\Delta_i = \mu_1 - \mu_i \leq \mu_1 - \mu_a + \tfrac{\Delta_a}{2} + \tfrac{\Delta_i}{2} \Rightarrow \Delta_i \leq 3\Delta_a. \tag{9}$$

Therefore, given event $E(T)$, the expected regret due to play of a saturated arm in one of the $V_j^{\ell,a}$ steps is at most $3\Delta_a$, and thus, the expected regret *due to playing saturated arms* in interval I_j is

$$\mathbb{E}[\mathcal{R}^s(I_j)|E(T)] \leq \mathbb{E}[\sum_{\ell=1}^{\gamma_j+1} \sum_{a=2}^{N} (3\Delta_a) V_j^{\ell,a}]. \tag{10}$$

Recall that $V_j^{\ell,a}$ denotes the number of steps before either event $M(t)$ happens or some arm other than a becomes the best saturated arm. Now, given event $E(T)$, while a is the best saturated arm, the event $M'(t)$, defined below, implies event $M(t)$,

$$M'(t) : \theta_1(t) > \mu_a + \Delta_a/2. \tag{11}$$

More precisely, if a is the best saturated arm, then $\mu_a + \Delta_a/2 = \max_{i \in C(t)} \mu_i + \Delta_i/2$, and if $E(T)$ holds, then $\max_{i \in C(t)} \mu_i + \Delta_i/2 \geq \max_{i \in C(t)} \theta_i(t)$, therefore $M'(t) \Rightarrow \theta_1(t) > \max_{i \in C(t)} \theta_i(t) \equiv M(t)$.

This means that given event $E(T)$, and given $s(j) = s$, $V_j^{\ell,a}$ is stochastically dominated by geometric random variable $X(j, s, \mu_a + \Delta_a/2)$ (which was defined as the number of trials until an independent sample from $\text{Beta}(s+1, j-s+1)$ distribution exceeds $\mu_a + \Delta_a/2$):

$$\mathbb{E}[V_j^{\ell,a}|s(j), E(T)] \leq \mathbb{E}[X(j, s(j), \mu_a + \frac{\Delta_a}{2})|s(j)]. \tag{12}$$

This gives

$$\begin{aligned}
\mathbb{E}[\mathcal{R}^s(I_j)|E(T)] &\leq \mathbb{E}[\sum_{\ell=1}^{\gamma_j+1}\sum_{a=2}^{N}(3\Delta_a)V_j^{\ell,a}|E(T)] \\
&= \mathbb{E}[\sum_{\ell=1}^{\gamma_j+1}\sum_{a=2}^{N}(3\Delta_a)\mathbb{E}[V_j^{\ell,a}|s(j),E(T)]] \\
&\leq \mathbb{E}[(\gamma_j+1)\sum_{a=2}^{N}(3\Delta_a)\min\{\mathbb{E}[X(j,s(j),\mu_a+\frac{\Delta_a}{2})|s(j)],T\} \quad \text{(using (12))} \qquad (13)
\end{aligned}$$

Now, the total expected regret due to saturated arms, given $E(T)$, is

$$\begin{aligned}
\mathbb{E}[\mathcal{R}^s(T)|E(T)] &= \sum_{j=0}^{T-1}\mathbb{E}[\mathcal{R}^s(I_j)|E(T)] \\
&\leq \sum_{j=0}^{T-1}\mathbb{E}[\gamma_j\sum_a(3\Delta_a)\mathbb{E}[X(j,s(j),\mu_a+\frac{\Delta_a}{2})|s(j)]] \\
&\qquad + \sum_{j=0}^{T-1}\mathbb{E}[\sum_a(3\Delta_a)\min\{\mathbb{E}[X(j,s(j),\mu_a+\frac{\Delta_a}{2})|s(j)],T\}] \\
&= \sum_{j=0}^{T-1}\mathbb{E}\left[\gamma_j\sum_a(3\Delta_a)\left(\frac{1}{F_{j+1,y_a}(s(j))}-1\right)\right] \\
&\qquad + \sum_j\mathbb{E}[\sum_a(3\Delta_a)\min\{\mathbb{E}[X(j,s(j),\mu_a+\frac{\Delta_a}{2})|s(j)],T\}], \qquad (14)
\end{aligned}$$

where $y_a = \mu_a + \Delta_a/2$. The last equality follows from Lemma 1, which gives the expression for the expected value of random variable $X(j,s,y)$. In this section, we abbreviate the symbol $F^B_{n,p}$ for the cdf of Binomial(n,p) distribution to $F_{n,p}$.

Recall that γ_j denotes the number of occurrences of event $M(t)$ in interval I_j, i.e. the number of times in interval I_j, $\theta_1(t)$ was greater than $\theta_i(t)$ of all saturated arms $i \in C(t)$, and still the first arm was not played. The only reason the first arm would not be played at a time t despite of $\theta_1(t) > \max_{i\in C(t)}\theta_i(t)$, is that some unsaturated arm u with highest $\theta_u(t)$ was played instead. And, since an unsaturated arm u can be played for at most L_u times before it becomes saturated, the random variables γ_j always satisfy

$$\sum_{j=0}^{T-1}\gamma_j \leq \sum_{u=2}^{N} L_u = \sum_{u=2}^{N}\frac{16\ln T}{\Delta_u^2}. \qquad (15)$$

Therefore, for the first term on the right hand side of (14) we have

$$
\begin{aligned}
&\sum_{j=0}^{T-1} \mathbb{E}[\gamma_j \sum_a \frac{3\Delta_a}{F_{j+1,y_a}(s(j))}] \\
= & \sum_a \sum_{j=0}^{T-1} \mathbb{E}[\gamma_j \frac{3\Delta_a}{F_{j+1,y_a}(s(j))}] \\
\leq & \sum_a (\sum_u L_u) \mathbb{E}[\max_j \frac{3\Delta_a}{F_{j+1,y_a}(s(j))}] \\
\leq & (\sum_u L_u) \sum_a \mathbb{E}[\frac{3\Delta_a}{F_{j_a^*+1,y_a}(s(j_a^*))} \cdot I(s(j_a^*) \leq \lfloor y_a j_a^* \rfloor) + \frac{3\Delta_a}{F_{j_a^*+1,y_a}(s(j_a^*))} \cdot I(s(j_a^*) \geq \lceil y_a j_a^* \rceil)] \\
\leq & (\sum_u L_u) \sum_a \mathbb{E}[\frac{3\Delta_a}{F_{j_a^*+1,y_a}(s(j_a^*))} \cdot I(s(j_a^*) \leq \lfloor y_a j_a^* \rfloor)] + (\sum_u L_u) \cdot \frac{6\Delta_a}{1-y_a}, \qquad (16)
\end{aligned}
$$

where we define random variable j_a^* as

$$
j_a^* = \arg \max_{j \in \{0,\ldots,T-1\}} \frac{1}{F_{j+1,y_a}(s(j))}.
$$

(Note that j_a^* is completely specified by the random sequence $s(1), s(2), \ldots.$) The last inequality follows from the fact that $F_{j+1,y}(s) \geq (1-y)F_{j,y}(s)$, and that for $s \geq \lceil yj \rceil$, $F_{j,y}(s) \geq 1/2$ (Fact 2). Now, for the first term in above,

$$
\begin{aligned}
\mathbb{E}[\frac{1}{F_{j_a^*+1,y_a}(s(j_a^*))} \cdot I(s(j_a^*) \leq \lfloor y_a j_a^* \rfloor)] \quad & \leq \quad \sum_j \mathbb{E}[\frac{1}{F_{j+1,y_a}(s(j))} \cdot I(s(j) \leq \lfloor y_a j \rfloor)] \\
& = \quad \sum_j \sum_{s=0}^{\lfloor y_a j \rfloor} \frac{f_{j,\mu_1}(s)}{F_{j+1,y_a}(s)} \leq \sum_j \frac{\mu_1}{\Delta_a'} e^{-D_a j} \leq \frac{16}{\Delta_a^3}, \qquad (17)
\end{aligned}
$$

where $\Delta_a' = \mu_1 - y_a = \Delta_a/2$, D_a is the KL-divergence between Bernoulli distributions with parameters μ_1 and y_a. The penultimate inequality follows using (3) in the proof of Lemma 3, with $\Delta' = \Delta_a'$, and $D = D_a$. The last inequality uses the geometric series sum (note that $D_a \geq 0$ by Gibbs' inequality),

$$
\textstyle\sum_j e^{-D_a j} \leq \frac{1}{1-e^{-D_a}} \leq \min\{\frac{2}{D_a}, \frac{e}{e-1}\} \leq \frac{2}{\min\{D_a,1\}} \leq \frac{2}{\Delta_a'^2} = \frac{8}{\Delta_a^2}.
$$

Substituting the bound from Equation (17) in Equation (16),

$$
\textstyle\sum_{j=0}^{T-1} \mathbb{E}\left[\gamma_j \sum_a \frac{3\Delta_a}{F_{j+1,y_a}(s(j))}\right] \leq (\sum_u L_u) \sum_a (\frac{48}{\Delta_a^2} + \frac{6\Delta_a}{\Delta_a'}) = (\sum_u L_u) \sum_a (\frac{48}{\Delta_a^2} + 12). \qquad (18)
$$

For the second term on the right hand side of (14), we use Lemma 3 while substituting y with

$y_a = \mu_a + \frac{\Delta_a}{2}$ and Δ' with $\mu_1 - y_a = \frac{\Delta_a}{2}$, to obtain that

$$\begin{aligned}
\sum_{j=0}^{T-1} \sum_a (3\Delta_a) \mathbb{E}\left[\min\{\mathbb{E}[X(j, s(j), \mu_a + \frac{\Delta_a}{2})|s(j)], T\} \right] \\
\leq \quad \sum_a (3\Delta_a) \sum_{j=0}^{\frac{16(\ln T)}{\Delta_a{}^2}-1} \left(1 + \frac{2}{1-y_a}\right) + \sum_{j \geq \frac{16(\ln T)}{\Delta_a{}^2}}^{T} (3\Delta_a)\frac{16}{T} \\
\leq \quad \sum_a \frac{48\ln T}{\Delta_a} + \frac{192}{\Delta_a^2} + 48\Delta_a. \qquad (19)
\end{aligned}$$

Substituting results from Equation (18) and (19) in Equation (14),

$$\begin{aligned}
\mathbb{E}[\mathcal{R}^s(T)|E(T)] &\leq (\sum_u L_u) \sum_a (\frac{48}{\Delta_a^2} + 12) + \sum_a (\frac{48 \ln T}{\Delta_a} + \frac{192}{\Delta_a^2} + 48\Delta_a) \\
&\leq 768(\ln T)(\sum_i \frac{1}{\Delta_i^2})^2 + 192(\ln T)\sum_i \frac{1}{\Delta_i^2} + 48(\ln T)\sum_a \frac{1}{\Delta_a} + 192\sum_a \frac{1}{\Delta_a^2} + 48(N-1).
\end{aligned}$$

Now, using the result that $\Pr(\overline{E(T)}) \leq 4(N-1)/T$ (by Lemma 4), we can bound the total regret due to playing saturated arms as

$$\begin{aligned}
\mathbb{E}[\mathcal{R}^s(T)] &\leq \mathbb{E}[\mathcal{R}^s(T)|E(T)] + T \cdot \Pr(\overline{E(T)}) \\
&\leq 768(\ln T)(\sum_i \frac{1}{\Delta_i^2})^2 + 192(\ln T)\sum_i \frac{1}{\Delta_i^2} + 48(\ln T)\sum_a \frac{1}{\Delta_a} + 192\sum_a \frac{1}{\Delta_a^2} + 52(N-1).
\end{aligned}$$

Since an unsaturated arm u becomes saturated after L_u plays, regret due to unsaturated arms is at most

$$\mathbb{E}[\mathcal{R}^u(T)] \leq \sum_{u=2}^{N} L_u \Delta_u = 16(\ln T)\left(\sum_{u=2}^{N} \frac{1}{\Delta_u}\right).$$

Summing the regret due to saturated and unsaturated arms, we obtain the result of Theorem 2.

Conclusion. In this paper, we showed theoretical guarantees for Thompson Sampling close to other state of the art methods, like UCB. Our result is a first step in theoretical understanding of TS and there are several avenues to explore for the future work: There is a gap between our upper bounds and the lower bound of Lai–Robbins [7]. While it may be easy to improve the constant factors in our upper bounds by making the analysis more careful (but more complicated), it seems harder to improve the dependence on the Δ's. With further work, our technique in this paper can provide several extensions, including analysis of TS for bandits with more general distributions than Bernoulli, delayed and batched feedbacks, prior mismatch and posterior reshaping discussed in [2]. As mentioned before, empirically TS has been shown to have superior performance than other methods, especially for handling delayed feedback. A theoretical justification of this observation would require a tighter analysis of TS than what we have achieved here, and in addition, it would require lower bound on the regret of the other algorithms. TS has also been used for problems such as regularized logistic regression [2]. These multi-parameter settings lack theoretical analysis.

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256, 2002.

[2] O. Chapelle and L. Li. An Empirical Evaluation of Thompson Sampling. In *NIPS*, 2011.

[3] J. C. Gittins. *Multi-armed Bandit Allocation Indices.* Wiley Interscience Series in Systems and Optimization. John Wiley and Son, 1989.

[4] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-Scale Bayesian Click-Through rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *ICML*, pages 13–20, 2010.

[5] O.-C. Granmo. Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton. *International Journal of Intelligent Computing and Cybernetics (IJICC)*, 3(2):207–234, 2010.

[6] K. Jogdeo and S. M. Samuels. Monotone Convergence of Binomial Probabilities and A Generalization of Ramanujan's equation. *The Annals of Mathematical Statistics*, (4):1191–1195, 1968.

[7] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[8] B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. Technical report, Statistics Group, Department of Mathematics, University of Bristol.

[9] B. C. May and D. S. Leslie. Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical report, Statistics Group, Department of Mathematics, University of Bristol.

[10] S. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639–658, 2010.

[11] W. 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.

A Multiple optimal arms

Consider the N-armed bandit problem with $\mu^* = \max_i \mu_i$. We will show that adding another arm with expected reward μ^* can only decrease the expected regret of TS algorithm. Suppose that we added arm $N+1$ with expected reward μ^*. Consider the expected regret for the new bandit in time T, conditioned on the exact time steps among $1, \ldots, T$, on which arm $N+1$ is played by the algorithm. Since the arm $N+1$ has expected reward μ^*, there is no regret in these time steps. Now observe that in the remaining time steps, the algorithm behaves exactly as it would for the original bandit with N arms. Therefore, given that the $(N+1)^{th}$ arm is played x times, the expected regret in time T for the new bandit will be same as the expected regret in time $T-x$ for the original bandit. Let $\mathcal{R}^N(T)$ and $\mathcal{R}^{N+1}(T)$ denote the expected regret in time T for the original and new bandit, respectively. Then,

$$\begin{aligned}\mathbb{E}[\mathcal{R}^{N+1}(T)] = \mathbb{E}[\mathbb{E}[\mathcal{R}^{N+1}(T)|k_{N+1}(T)]] &= \mathbb{E}[\mathbb{E}[\mathcal{R}^N(T-k_{N+1}(T))|k_{N+1}(T)]] \\ &\leq \mathbb{E}[\mathbb{E}[\mathcal{R}^N(T)|k_{N+1}(T)]] = \mathbb{E}[\mathcal{R}^N(T)].\end{aligned}$$

This argument shows that the expected regret of Thompson Sampling for the N-armed bandit problem with r optimal arms is bounded by the expected regret of Thompson Sampling for the $(N-r+1)$-armed bandit problem obtained on removing (any) $r-1$ of the optimal arms.

B Facts used in the analysis

Fact 1.

$$F^{beta}_{\alpha,\beta}(y) = 1 - F^B_{\alpha+\beta-1,y}(\alpha-1),$$

for all positive integers α, β.

Proof. This fact is well-known (it's mentioned on Wikipedia) but we are not aware of a specific reference. Since the proof is easy and short we will present a proof here. The Wikipedia page also mentions that it can be proved using integration by parts. Here we provide a direct combinatorial proof which may be new.

One well-known way to generate a r.v. with cdf $F^{beta}_{\alpha,\beta}$ for integer α and β is the following: generate uniform in $[0,1]$ r.v.s $X_1, X_2, \ldots, X_{\alpha+\beta-1}$ independently. Let the values of these r.v. in sorted increasing order be denoted $X^\uparrow_1, X^\uparrow_2, \ldots, X^\uparrow_{\alpha+\beta-1}$. Then $X^\uparrow_\alpha$ has cdf $F^{beta}_{\alpha,\beta}$. Thus $F^{beta}_{\alpha,\beta}(y)$ is the probability that $X^\uparrow_\alpha \leq y$.

We now reinterpret this probability using the binomial distribution: The event $X^\uparrow_\alpha \leq y$ happens iff for at least α of the $X_1, \ldots, X_{\alpha+\beta-1}$ we have $X_i \leq y$. For each X_i we have $\Pr[X_i \leq y] = y$; thus the probability that for at most $\alpha - 1$ of the X_i's we have $X_i \leq y$ is $F^B_{\alpha+\beta-1,y}(\alpha-1)$. And so the probability that for at least α of the X_i's we have $X_i \leq y$ is $1 - F^B_{\alpha+\beta-1,y}(\alpha-1)$. □

The median of an integer-valued random variable X is an integer m such that $\Pr(X \leq m) \geq 1/2$ and $\Pr(X \geq m) \geq 1/2$. The following fact says that the median of the binomial distribution is close to its mean.

Fact 2 ([6])**.** *Median of the binomial distribution Binomial(n,p) is either $\lfloor np \rfloor$ or $\lceil np \rceil$.*

Fact 3 ([1])**.** (Chernoff–Hoeffding bound) *Let $X_1, ..., X_n$ be random variables with common range $[0,1]$ and such that $\mathbb{E}[X_t|X_1, ..., X_{t-1}] = \mu$. Let $S_n = X_1 + \ldots + X_n$. Then for all $a \geq 0$,*

$$\Pr(S_n \geq n\mu + a) \leq e^{-2a^2/n},$$

$$\Pr(S_n \leq n\mu - a) \leq e^{-2a^2/n}.$$

Lemma 5.

$$F^B_{n,p}(np - n\delta) \leq e^{-2n\delta^2}, \quad 1 - F^B_{n,p}(np + n\delta) \leq e^{-2n\delta^2}, \tag{20}$$

$$1 - F^B_{n+1,p}(np + n\delta) \leq \frac{e^{4\delta}}{e^{2n\delta^2}}. \tag{21}$$

Proof. The first result is a simple application of Chernoff–Hoeffding bounds from Lemma 3. For the second result, we observe that,

$$F^B_{n+1,p}(np+n\delta) = (1-p)F^B_{n,p}(np+n\delta) + pF^B_{n,p}(np+n\delta-1) \geq F^B_{n,p}(np+n\delta-1).$$

By Chernoff–Hoeffding bounds,

$$1 - F^B_{n,p}(np+\delta n-1) \leq e^{-2(\delta n-1)^2/n} = e^{-2(n^2\delta^2+1-2\delta n)/n} \leq e^{-2n\delta^2+4\delta} = \frac{e^{4\delta}}{e^{2n\delta^2}}.$$

□

C Proofs of Lemmas

C.1 Proof of Lemma 2

Proof. Recall that $k_2(t)$ denotes the number of plays of second arm before time t, and $S_2(t)$ denotes the number of successes among these $k_2(t)$ plays. Define $A(t)$ to be the event that $\frac{S_2(t)}{k_2(t)} \leq \mu_2 + \frac{\Delta}{4}$. By the Chernoff–Hoeffding bounds (Fact 3), at any time t such that $k_2(t) \geq L = 16(\ln T)/\Delta^2$,

$$\Pr(\overline{A(t)}) = \Pr\left(\frac{S_2(t)}{k_2(t)} > \mu_2 + \frac{\Delta}{4}\right) \leq e^{-2L\Delta^2/16} \leq e^{-2\ln T} = \frac{1}{T^2}.$$

Then, for every time t such that $k_2(t) \geq L$, we have that

$$\begin{aligned}
\Pr(\theta_2(t) > \mu_2 + \frac{\Delta}{2}) &\leq \Pr(\theta_2(t) > \mu_2 + \frac{\Delta}{2} | A(t)) + \Pr(\overline{A(t)}) \\
&\leq \Pr(\theta_2(t) > \frac{S_2(t)}{k_2(t)} - \frac{\Delta}{4} + \frac{\Delta}{2}) + \Pr(\overline{A(t)}) \\
&= F_{k_2(t)+1, \frac{S_2(t)}{k_2(t)} + \frac{\Delta}{4}}(S_2(t)) + \Pr(\overline{A(t)}) \\
&\leq F_{k_2(t), \frac{S_2(t)}{k_2(t)} + \frac{\Delta}{4}}(S_2(t)) + \Pr(\overline{A(t)}) \\
&\leq \exp\{-\frac{2\Delta^2 k_2(t)^2/16}{k_2(t)}\} + \frac{1}{T^2} \\
&\leq e^{-2L\Delta^2/16} + \frac{1}{T^2} = e^{-2\ln T} + \frac{1}{T^2} = \frac{2}{T^2}.
\end{aligned}$$

The third-last inequality follows from the observation that

$$F^B_{n+1,p}(r) = (1-p)F^B_{n,p}(r) + pF_{n,p}(r-1) \leq (1-p)F^B_{n,p}(r) + pF_{n,p}(r) = F^B_{n,p}(r).$$

And, the second-last inequality follows from Chernoff–Hoeffding bounds (refer to Fact 3 and Lemma 5).

Summing above over $t = 1, \ldots, T$, we get the result of the lemma. □

C.2 Proof of Lemma 4

Proof. Define $A_i(t)$ to be the event that $\frac{S_i(t)}{k_i(t)} \leq \mu_i + \frac{\Delta_i}{4}$. By the Chernoff–Hoeffding bounds (Fact 3), at any time t such that arm i has been played at least $L_i = 16(\ln T)/\Delta_i^2$ times before t,

$$1 - \Pr(A_i(t)) = \Pr\left(\frac{S_i(t)}{k_i(t)} > \mu_i + \frac{\Delta_i}{4}\right) \leq e^{-2L_i\Delta_i^2/16} \leq e^{-2\ln T} \leq \frac{1}{T^2}.$$

Then, for every time t after L_i plays of arm i,

$$\begin{aligned}
\Pr\left(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}\right) &\leq \Pr(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}|A_i(t)) + (1 - \Pr(A_i(t))) \\
&\leq \Pr(\theta_i(t) > \frac{S_i(t)}{k_i(t)} - \frac{\Delta_i}{4} + \frac{\Delta_i}{2}) + 1 - \Pr(A(t)) \\
&= F_{k_i(t)+1, \frac{S_i(t)}{k_i(t)} + \frac{\Delta_i}{4}}(S_i(t)) + 1 - \Pr(A_i(t)) \quad \text{(Fact 1)} \\
&\leq F_{k_i(t), \frac{S_i(t)}{k_i(t)} + \frac{\Delta_i}{4}}(S_i(t)) + 1 - \Pr(A_i(t)) \\
&\leq \exp\{-\frac{2\Delta_i^2 k_i(t)^2/16}{k_i(t)}\} + \frac{1}{T^2} \\
&\leq e^{-2L_i\Delta_i^2/16} + \frac{1}{T^2} = e^{-2\ln T} + \frac{1}{T^2} = \frac{2}{T^2}.
\end{aligned}$$

The third last inequality follows from the observation that

$$F^B_{n+1,p}(r) = (1-p)F^B_{n,p}(r) + pF_{n,p}(r-1) \leq (1-p)F^B_{n,p}(r) + pF_{n,p}(r) = F^B_{n,p}(r).$$

And, second last inequality follows from Chernoff–Hoeffding bounds (refer to Fact 3 and Lemma 5).

Similarly, we can derive that

$$\Pr(\theta_i(t) < \mu_i - \frac{\Delta_i}{2}) \leq \frac{2}{T^2},$$

Summing above over $t = 1, \ldots, T, i = 2, \ldots, N$, we get the result of the lemma. □