JMLR: Workshop and Conference Proceedings vol 23 (2012) 39.1–39.26 25th Annual Conference on Learning Theory

Analysis of Thompson Sampling for the Multi-armed Bandit Problem

Shipra Agrawal SHIPRA@MICROSOFT.COM
Microsoft Research India

Navin Goyal NAVINGO@MICROSOFT.COM
Microsoft Research India

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

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 stochastic two-armed bandit problem, the expected regret in time T is $O(\frac{\ln T}{\Delta} + \frac{1}{\Delta^3})$. And, for the stochastic N-armed 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.

Keywords: multi-armed bandit, Thompson Sampling, Bayesian algorithm, online learning

1. Introduction

Multi-armed bandit problem models the exploration/exploitation trade-off inherent in sequential decision problems. 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: the 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., Lai and Robbins (1985); Auer et al. (2002), and more recently Garivier and Cappé (2011), Maillard et al. (2011), Kaufmann et al. (2012)), which have good theoretical guarantees, and the algorithm by Gittins (1989), which gives optimal strategy under Bayesian setting with known priors and geometric time-discounted rewards. In one of the earliest works on stochastic bandit problems, Thompson (1933) proposed a natural randomized Bayesian algorithm to minimize regret. The basic idea is to assume a simple prior distribution on the parameters of the reward distribution of every arm, and at any time step, play an arm according to its posterior 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. We emphasize that although TS algorithm is a Bayesian approach, the description of the algorithm and our analysis apply to the prior-free stochastic multi-armed bandit model where parameters of the reward distribution of every arm are

fixed, though unknown (see Section 1.1). One could interpret the "assumed" Bayesian priors as the current knowledge of the algorithm about the arms. Thus, our regret bounds for Thompson Sampling are directly comparable to the regret bounds for UCB family of algorithms which are a frequentist approach to the same problem.

Recently, TS has attracted considerable attention. Several studies (e.g., Granmo (2010); Scott (2010); Chapelle and Li (2011); May and Leslie (2011)) have empirically demonstrated the efficacy of Thompson Sampling: Scott (2010) provides a detailed discussion of probability matching techniques in many general settings along with favorable empirical comparisons with other techniques. Chapelle and Li (2011) demonstrate that empirically TS achieves regret comparable to the lower bound of Lai and Robbins (1985); 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 (delayed feedback means that the result of a play of an arm may become available only after some time delay, but we are required to make immediate decisions for which arm to play next) 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 (Graepel et al. (2010)) for CTR prediction of search ads on Bing uses the idea of Thompson Sampling.

It has been suggested (Chapelle and Li (2011)) 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 in Granmo (2010); May et al. (2011) provide weak guarantees, namely, a bound of $o(T)$ on expected regret in time T. 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 Lai and Robbins (1985). 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

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

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

1.2. Thompson Sampling

For simplicity of discussion, we first provide the details of Thompson Sampling 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 (2011). 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(α, β), 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(α, β) around the mean. Beta distribution is useful for Bernoulli rewards because if the prior is a Beta(α, β) distribution, then after observing a Bernoulli trial, the posterior distribution is simply Beta$(\alpha+1, \beta)$ or Beta$(\alpha, \beta+1)$, depending on whether the trial resulted in a success or failure, respectively.

The Thompson Sampling algorithm initially 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 updates the distribution on μ_i as Beta$(S_i(t) + 1, F_i(t) + 1)$. The algorithm then samples from these posterior distributions 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

For each arm $i = 1, \ldots, N$ set $S_i = 0, F_i = 0$.
foreach $t = 1, 2, \ldots,$ **do**
  For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ from the Beta$(S_i + 1, F_i + 1)$ distribution.
  Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward r_t.
  If $r = 1$, then $S_{i(t)} = S_{i(t)} + 1$, else $F_{i(t)} = F_{i(t)} + 1$.
end

We adapt the Bernoulli Thompson sampling algorithm to the general stochastic bandits case, i.e. when the rewards for arm i are generated from an arbitrary unknown distribution with support $[0, 1]$ and mean μ_i, in a way that allows us to reuse our analysis of the Bernoulli case. To our knowledge, this adaptation is new. We modify TS so that after observing the reward $\tilde{r}_t \in [0, 1]$ at time t, it performs a Bernoulli trial with success probability $\tilde{r_t}$. Let random variable r_t denote the outcome of this Bernoulli trial, and let $\{S_i(t), F_i(t)\}$ denote the number of successes and failures in the Bernoulli trials until time t. The remaining algorithm is the same as for Bernoulli bandits. Algorithm 2 gives the precise description of this algorithm.

We observe that the probability of observing a success (i.e., $r_t = 1$) in the Bernoulli trial after playing an arm i in the new generalized algorithm is equal to the mean reward μ_i. Let f_i denote the (unknown) pdf of reward distribution for arm i. Then, on playing arm i,

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

Algorithm 2 Thompson Sampling for general stochastic bandits

For each arm $i = 1, \ldots, N$ set $S_i(1) = 0, F_i(1) = 0$.
foreach $t = 1, 2, \ldots,$ **do**
  For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ from the Beta$(S_i + 1, F_i + 1)$ distribution.
  Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward $\tilde{r}_t$.
  Perform a Bernoulli trial with success probability $\tilde{r}_t$ and observe output r_t.
  If $r_t = 1$, then $S_{i(t)} = S_{i(t)} + 1$, else $F_{i(t)} = F_{i(t)} + 1$.
end

Thus, the probability of observing $r_t = 1$ 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. Therefore, the analysis of TS for Bernoulli setting is applicable to this modified TS for the general setting. This allows us to replace, for the purpose of analysis, the problem with general stochastic bandits with Bernoulli bandits with the same means. We remark that instead of using r_t, we could consider more direct and natural updates of type Beta(α_i, β_i) to Beta$(\alpha_i + \tilde{r_t}, \beta_i + 1 - \tilde{r_t})$. However, we do not know how to analyze this because of our essential use of Fact 1, which requires α_i, β_i to be integral.

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*

$$\mathbb{E}[\mathcal{R}(T)] = O\left(\frac{\ln T}{\Delta} + \frac{1}{\Delta^3}\right)$$

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

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

$$\mathbb{E}[\mathcal{R}(T)] \leq O\left(\left(\sum_{a=2}^{N} \frac{1}{\Delta_a^2}\right)^2 \ln T\right)$$

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

Remark 3 *For the N-armed bandit problem, we can obtain an alternate bound of*

$$\mathbb{E}[\mathcal{R}(T)] \leq O\left(\frac{\Delta_{max}}{\Delta_{min}^3}\left(\sum_{a=2}^{N} \frac{1}{\Delta_a^2}\right) \ln T\right)$$

by slight modification to the proof. The above bound has a better dependence on N than in Theorem 2, but worse dependence on Δ_is. Here $\Delta_{min} = \min_{i \neq 1} \Delta_i$,$\Delta_{max} = \max_{i \neq 1} \Delta_i$.

In interest of readability, we used big-Oh notation [1] to state our results. The exact constants are provided in the proofs of the above theorems. Let us contrast our bounds with the previous work. Lai and Robbins (1985) proved the following lower bound on regret of any bandit algorithm:

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

where D denotes the KL divergence. They also gave algorithms asymptotically achieving this guarantee, though unfortunately their algorithms are not efficient. Auer et al. (2002) gave the UCB1 algorithm, which is efficient and achieves the following bound:

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

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. (2002). For the N-armed setting, the exponent of Δ's in our bound is basically 4 compared to the exponent 1 for UCB1.

More recently, Kaufmann et al. (2012) gave Bayes-UCB algorithm which achieves the lower bound of Lai and Robbins (1985) for Bernoulli rewards. Bayes-UCB is a UCB like algorithm, where the upper confidence bounds are based on the quantiles of Beta posterior distributions. Interestingly, these upper confidence bounds turn out to be similar to those used by algorithms in Garivier and Cappé (2011) and Maillard et al. (2011). Bayes-UCB can be seen as an hybrid of TS and UCB. However, the general structure of the arguments used in Kaufmann et al. (2012) is similar to Auer et al. (2002); for the analysis of Thompson Sampling we need to deal with additional difficulties, as discussed in the next section.

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 (for example, in Auer et al. (2002)): once the second arm has been

1. For any two functions $f(n), g(n)$, $f(n) = O(g(n))$ if there exist two constants n_0 and c such that for all $n \geq n_0$, $f(n) \leq cg(n)$.

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. (2002), 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 $L = 24(\ln T)/\Delta^2$ times. The expected regret before this event is bounded by $24(\ln T)/\Delta$ 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 $\mathbb{E}\left[X(j, s(j), \mu_2) \,|\, s(j)\right] = 1/\Pr[\theta_1 > \mu_2] - 1$. 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 $\mathbb{E}\left[\mathbb{E}\left[X(j, s(j), \mu_2) \,|\, s(j)\right]\right]$. However, for small j we do not have such concentration, and it requires a delicate computation to get a bound on $\mathbb{E}\left[\mathbb{E}\left[X(j, s(j), \mu_2) \,|\, s(j)\right]\right]$. 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 ($L_a = 24(\ln T)/\Delta_a^2$) 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 closely approximated 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 arms 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 $L = 24(\ln T)/\Delta^2$ plays of the second arm. Let random variable t_j denote the time step at which the j^{th} play of the first arm happens (we define $t_0 = 0$). Also, let random variable $Y_j = t_{j+1} - t_j - 1$ 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

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

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. Also, let $F^{B}_{n,p}$ denote the cdf of the *binomial* distribution with parameters (n, p).

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

Lemma 4 *For all non-negative integers $j, s \le j$, and for all $y \in [0,1]$,*

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

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, $\mathbb{E}\left[X(j,s,y)\right] = \frac{1}{1-F^{beta}_{s+1,j-s+1}(y)} - 1$ (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. ■

Recall that Y_j was defined as the number of steps before $\theta_1(t) > \theta_2(t)$ happens for the first time after the j^{th} play of the first arm. Now, consider the number of steps before $\theta_1(t) > \mu_2 + \frac{\Delta}{2}$ happens for the first time after the j^{th} play of the first arm. Given $s(j)$, this has the same distribution as $X(j, s(j), \mu_2 + \frac{\Delta}{2})$. However, Y_j can be larger than this number if (and only if) at some time step t between t_j and t_{j+1}, $\theta_2(t) > \mu_2 + \frac{\Delta}{2}$. In that case we use the fact that Y_j is always bounded by T. Thus, for any $j \ge j_0$, we can bound $\mathbb{E}[Y_j]$ as,

$$\mathbb{E}[Y_j] \le \mathbb{E}[\min\{X(j,s(j),\mu_2+\tfrac{\Delta}{2}), T\}] + \mathbb{E}[\textstyle\sum_{t=t_j+1}^{t_{j+1}-1} T \cdot I(\theta_2(t) > \mu_2 + \frac{\Delta}{2})].$$

Here notation $I(E)$ is the indicator for event E, i.e., its value is 1 if event E happens and 0 otherwise. In the first term of RHS, the expectation is over distribution of $s(j)$ as well as over the distribution of the geometric variable $X(j, s(j), \mu_2 + \frac{\Delta}{2})$. Since we are interested only in $j \ge j_0$, we will instead use the similarly obtained bound on $\mathbb{E}[Y_j \cdot I(j \ge j_0)]$,

$$\mathbb{E}[Y_j \cdot I(j \ge j_0)] \le \mathbb{E}[\min\{X(j,s(j),\mu_2+\tfrac{\Delta}{2}), T\}] + \mathbb{E}[\textstyle\sum_{t=t_j+1}^{t_{j+1}-1} T \cdot I(\theta_2(t) > \mu_2 + \frac{\Delta}{2}) \cdot I(j \ge j_0)].$$

This gives,

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{j=j_0}^{T-1} Y_j] &\le \textstyle\sum_{j=0}^{T-1} \mathbb{E}[\min\{X(j,s(j),\mu_2+\frac{\Delta}{2}), T\}] + T \cdot \sum_{j=0}^{T-1} \mathbb{E}[\sum_{t=t_j+1}^{t_{j+1}-1} I(\theta_2(t) > \mu_2 + \tfrac{\Delta}{2}, j \ge j_0)] \\ &\le \textstyle\sum_{j=0}^{T-1} \mathbb{E}[\min\{X(j,s(j),\mu_2+\frac{\Delta}{2}), T\}] + T \cdot \sum_{t=1}^{T} \Pr(\theta_2(t) > \mu_2 + \tfrac{\Delta}{2}, k_2(t) \ge L).\end{aligned}$$

The last inequality holds because for any $t \in [t_j + 1, t_{j+1} - 1], j \ge j_0$, by definition $k_2(t) \ge L$. We denote the event $\{\theta_2(t) \le \mu_2 + \frac{\Delta}{2}$ or $k_2(t) < L\}$ by $E_2(t)$. In words, this is the event that if sufficient number of plays of second arm have happened until time t, then $\theta_2(t)$ is not much larger than μ_2; intuitively, we expect this event to be a high probability event as we will show. $\overline{E_2(t)}$ is the event $\{\theta_2(t) > \mu_2 + \frac{\Delta}{2}$ and $k_2(t) \ge L\}$ used in the above equation. Next, we bound $\Pr(E_2(t))$ and $\mathbb{E}[\min\{X(j,s(j),\mu_2+\frac{\Delta}{2}), T\}]$.

Lemma 5 $\forall t, \quad \Pr(E_2(t)) \ge 1 - \frac{2}{T^2}$.

Proof Refer to Appendix C.1. ■

Lemma 6 *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. $D = y \ln \frac{y}{\mu_1} + (1-y) \ln \frac{1-y}{1-\mu_1}$.*

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

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

Proof The complete proof of this lemma is included in Appendix C.2; here we provide some high level ideas.

Using Lemma 4, the expected value of $X(j, s(j), y)$ for any given $s(j)$,

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

For large j, i.e., $j \geq 4(\ln T)/\Delta'^2$, we use Chernoff–Hoeffding bounds to argue that with probability at least $(1 - \frac{8}{T^2})$, $s(j)$ will be greater than $\mu_1 j - \Delta' j/2$. And, for $s(j) \geq \mu_1 j - \Delta' j/2 = yj + \Delta' j/2$, we can show that the probability $F^B_{j+1,y}(s(j))$ will be at least $1 - \frac{8}{T^2}$, again using Chernoff–Hoeffding bounds. These observations allow us to derive that $\mathbb{E}\left[\mathbb{E}\left[\min\{X(j, s(j), y), T\}\right]\right] \leq \frac{16}{T}$, for $j \geq 4(\ln T)/\Delta'^2$.

For small j, the argument is more delicate. In this case, $s(j)$ could be small with a significant probability. More precisely, $s(j)$ could take a value s smaller than yj with binomial probability $f^B_{j,\mu_1}(s)$. For such s, we use the lower bound $F^B_{j+1,y}(s) \geq (1-y)F^B_{j,y}(s) + yF^B_{j,y}(s-1) \geq (1-y)F^B_{j,y}(s) \geq (1-y)f^B_{j,y}(s)$, and then bound the ratio $f^B_{j,\mu_1}(s)/f^B_{j,y}(s)$ in terms of Δ', R and KL-divergence D. For $s(j) = s \geq \lceil yj \rceil$, we use the observation that since $\lceil yj \rceil$ is greater than or equal to the median of Binomial(j, y) (see Jogdeo and Samuels (1968)), we have $F^B_{j,y}(s) \geq 1/2$. After some algebraic manipulations, we get the result of the lemma. ■

Using Lemma 5, and Lemma 6 for $y = \mu_2 + \Delta/2$, and $\Delta' = \Delta/2$, we can bound the expected number of plays of the second arm as:

$$\begin{aligned}
\mathbb{E}[k_2(T)] &= L + \mathbb{E}\left[\textstyle\sum_{j=j_0}^{T-1} Y_j\right] \\
&\leq L + \textstyle\sum_{j=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[\min\{X(j, s(j), \mu_2 + \frac{\Delta}{2}), T\} \,\middle|\, s(j)\right]\right] + \sum_{t=1}^{T} T \cdot \Pr(\overline{E_2(t)}) \\
&\leq L + \tfrac{4\ln T}{\Delta'^2} + \textstyle\sum_{j=0}^{4(\ln T)/\Delta'^2 - 1} \frac{\mu_1}{\Delta'} e^{-Dj} + \left(\frac{y}{D}\ln R\right)\frac{2}{1-y} + \sum_{j=\frac{y}{D}\ln R}^{4(\ln T)/\Delta'^2 - 1} \frac{R^y e^{-Dj}}{1-y} + \frac{16}{T}\cdot T + 2 \\
&\leq \tfrac{40\ln T}{\Delta^2} + \tfrac{48}{\Delta^4} + 18,
\end{aligned} \tag{1}$$

where the last inequality is obtained after some algebraic manipulations; details are provided in Appendix C.3.

This gives a regret bound of

$$\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}\left[\Delta \cdot\ k_2(T)\right] \leq \left(\frac{40\ln T}{\Delta} + \frac{48}{\Delta^3} + 18\Delta\right).$$

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.

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 $L_i := \frac{24\ln T}{\Delta_i^2}$ times before time t. We bound the regret due to playing unsaturated and saturated suboptimal arms separately. The former is easily bounded as we will see; most of the work is

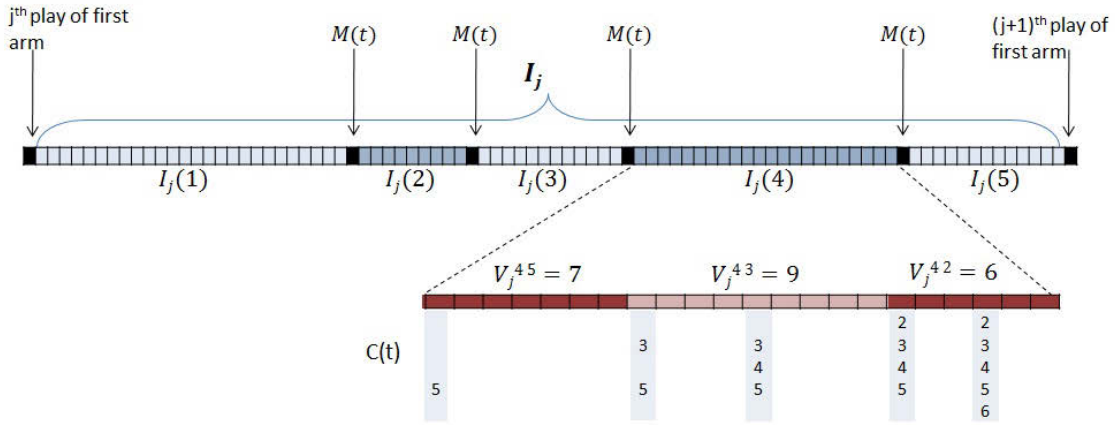

Figure 1: Interval I_j

in bounding the latter. For this, we bound the number of plays of saturated arms between two consecutive plays of the first arm.

In the following, by 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 $(j+1)^{th}$ plays of the first arm. We say that event $M(t)$ holds at time t, if $\theta_1(t)$ exceeds $\mu_i + \frac{\Delta_i}{2}$ of all the saturated arms, i.e.,

$$M(t) : \theta_1(t) > \max_{i \in C(t)} \mu_i + \frac{\Delta_i}{2}. \tag{2}$$

For t such that $C(t)$ is empty, we define $M(t)$ to hold trivially.

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

$$\gamma_j = |\{t \in I_j : M(t) = 1\}|. \tag{3}$$

Events $M(t)$ divide I_j into sub-intervals in a natural way: For $\ell = 2$ to γ_j, let r.v. $I_j(\ell)$ denote the sub-interval of I_j between the $(\ell-1)^{th}$ and ℓ^{th} occurrences of event $M(t)$ in I_j (excluding the time steps in which the event $M(t)$ occurs). We also define $I_j(1)$ and $I_j(\gamma_j + 1)$: If $\gamma_j > 0$ then $I_j(1)$ denotes the sub-interval in I_j before the first occurrence of event $M(t)$ in I_j; and $I_j(\gamma_j + 1)$ denotes the sub-interval in I_j after the last occurrence of event $M(t)$ in I_j. For $\gamma_j = 0$ we have $I_j(1) = I_j$.

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

Let us define event $E(t)$ as

$$E(t) : \ \{\theta_i(t) \in [\mu_i - \Delta_i/2, \mu_i + \Delta_i/2], \forall i \in C(t)\}.$$

In words, $E(t)$ denotes the event that all saturated arms have $\theta_i(t)$ tightly concentrated around their means. Intuitively, from the definition of saturated arms, $E(t)$ should hold with high probability; we prove this in the lemma below.

Lemma 7 *For all t,* $\Pr(E(t)) \geq 1 - \frac{4(N-1)}{T^2}$.
Also, for all t, j, and $s \leq j$, $\Pr(E(t) \mid s(j) = s) \geq 1 - \frac{4(N-1)}{T^2}$.

Proof Refer to Appendix C.4. ■

The stronger bound given by the second statement of lemma above will be useful later in the proof.

Observe that since a saturated arm i can be played at a step t only if $\theta_i(t)$ is greater than $\theta_1(t)$, the saturated arm i can be played at a time step t where $M(t)$ holds only if $\theta_i(t) > \mu_i + \Delta_i/2$. Thus, unless the high probability event $E(t)$ is violated, $M(t)$ denotes a play of an unsaturated arm at time t, and γ_j essentially denotes the number of plays of unsaturated arms in interval I_j. And, the number of plays of saturated arms in interval I_j is at most

$$\sum_{\ell=1}^{\gamma_j+1} |I_j(\ell)| + \sum_{t\in I_j} I(\overline{E(t)}).$$

We are interested in bounding regret due to playing saturated arms, which depends not only on the number of plays, but also on *which* saturated arm is played at each time step. Let $V_j^{\ell,a}$ denote the number of steps in $I_j(\ell)$, for which a is the best saturated arm, i.e.

$$V_j^{\ell,a} = |\{t \in I_j(\ell) : \mu_a = \max_{i\in C(t)} \mu_i\}|, \tag{4}$$

(resolve the ties for best saturated arm using an arbitrary, but fixed, ordering on arms). 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.

Recall that $M(t)$ holds trivially for all t such that $C(t)$ is empty. Therefore, there is at least one saturated arm at all $t \in I_j(\ell)$, and hence $V_j^{\ell,a}, a = 2, \ldots, N$ are well defined and cover the interval $I_j(\ell)$,

$$|I_j(\ell)| = \sum_{a=2}^{N} V_j^{\ell,a}.$$

Next, we will show that the regret due to playing any saturated arm at a time step t in one of the $V_j^{\ell,a}$ steps is at most $3\Delta_a + I(\overline{E(t)})$. The idea is that if all saturated arms have their $\theta_i(t)$ tightly concentrated around their means μ_i, then 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. That is, if a saturated arm i is played at a time t among one of the $V_j^{\ell,a}$ steps, then, either $E(t)$ is violated, i.e. $\theta_{i'}(t)$ for some saturated arm i' is not close to its mean, or

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

which implies that

$$\Delta_i = \mu_1 - \mu_i \leq \mu_1 - \mu_a + \tfrac{\Delta_a}{2} + \tfrac{\Delta_i}{2} \Rightarrow \Delta_i \leq 3\Delta_a. \tag{5}$$

Therefore, regret due to play of a saturated arm at a time t in one of the $V_j^{\ell,a}$ steps is at most $3\Delta_a + I(\overline{E(t)})$. With slight abuse of notation let us use $t \in V_j^{\ell,a}$ to indicate that t is one of the $V_j^{\ell,a}$ steps in $I_j(\ell)$. Then, the expected regret *due to playing saturated arms* in interval I_j is bounded as

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}^s(I_j)\right] &\leq \mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_{a=2}^{N}\sum_{t\in V_j^{\ell,a}}(3\Delta_a + I(\overline{E(t)}))\right] + \sum_{t\in I_j} I(\overline{E(t)}). \\
&= \mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_{a=2}^{N} 3\Delta_a V_j^{\ell,a}\right] + 2\mathbb{E}\left[\sum_{t\in I_j} I(\overline{E(t)})\right].
\end{aligned} \tag{6}$$

The second term in above will be bounded using Lemma 7. For bounding the first term, we establish the following lemma.

Lemma 8 *For all j,*

$$\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_a V_j^{\ell,a}\Delta_a\right] \le \mathbb{E}\left[\mathbb{E}\left[(\gamma_j+1)\,|\,s(j)\right]\sum_{a=2}^{N}\Delta_a\mathbb{E}\left[\min\{X(j,s(j),\mu_a+\tfrac{\Delta_a}{2}),T\}\,|\,s(j)\right]\right] \tag{7}$$

Proof The key observation used in proving this lemma is that given a fixed value of $s(j)=s$, the random variable $V_j^{\ell,a}$ is stochastically dominated by random variable $X(j,s,\mu_a+\frac{\Delta_a}{2})$ (defined earlier as a geometric variable denoting the number of trials before an independent sample from Beta$(s+1,j-s+1)$ distribution exceeds $\mu_a+\frac{\Delta_a}{2}$). A technical difficulty in deriving the inequality above is that the random variables γ_j and $V_j^{\ell,a}$ are not independent in general (both depend on the values taken by $\{\theta_i(t)\}$ over the interval). This issue is handled through careful conditioning of the random variables on history. The details of the proof are provided in Appendix C.5. ■

Next we illustrate the main ideas of the remaining proof by proving a weaker bound of $\left(\sum_i \frac{\log T}{\Delta_i^2}\right)^2$ on the expected regret. The proof of the bound $(\log T)\left(\sum_i \frac{1}{\Delta_i^2}\right)^2$ of Theorem 2 requires a slightly more careful analysis of this part, the complete details are given in Appendix D.

Consider the regret due to playing saturated arms until $\sum_{i=2}^N L_i$ plays of the first arm. After these many plays, the first arm will be concentrated enough so that the probability of playing any saturated arm (and hence the regret) will be very small. Now, using Lemma 8, the regret contributed by the first term in (6) can be loosely bounded by

$$\begin{aligned} & 3\mathbb{E}\left[\sum_{j=0}^{\sum_i L_i}\mathbb{E}\left[(\gamma_j+1)\,|\,s(j)\right]\sum_a\Delta_a\mathbb{E}\left[\min\{X(j,s(j),y_a),T\}\,|\,s(j)\right]\right] \\ \le\ & 3\mathbb{E}\left[\left(\sum_{j=0}^{\sum_i L_i}\mathbb{E}\left[(\gamma_j+1)\,|\,s(j)\right]\right)\left(\sum_{j=0}^{\sum_i L_i}\sum_a\Delta_a\mathbb{E}\left[\min\{X(j,s(j),y_a),T\}\,|\,s(j)\right]\right)\right]. \end{aligned}$$

Recall that γ_j is (approximately) the total number of plays of unsaturated arms in interval I_j. Therefore, the first term in the product above is bounded by the total number of plays of unsaturated arms, i.e. $O(\sum_{i=2}^N L_i)$. For the second term, using Lemma 6, we observe that $\mathbb{E}\left[\mathbb{E}\left[\min\{X(j,s(j),y_a),T\}\,|\,s(j)\right]\right]$ is bounded by $O(\frac{1}{\Delta_a})$. Therefore, the second term is bounded by $O(\sum_{i=2}^N L_i)$ as well. This gives a bound of $O((\sum_i L_i)^2) = O(\left(\sum_i \frac{\log T}{\Delta_i^2}\right)^2)$ on the above, and thus on the contribution of the first term of (6) towards the regret. The total contribution of the second term in Equation (6) can be bounded by a constant using Lemma 7.

Since an unsaturated arm u becomes saturated after L_u plays, regret due to unsaturated arms is at most $\sum_{u=2}^N L_u\Delta_u = 24(\ln T)\left(\sum_{u=2}^N \frac{1}{\Delta_u}\right)$. Summing the regret due to saturated and unsaturated arms, we obtain the weaker bound of $O((\sum_i \frac{\log T}{\Delta_i^2})^2)$ on regret. For details of the proof of the tighter bound of Theorem 2, see appendix D.

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. With further work, we hope that our techniques in this paper will be useful in providing several extensions, including a tighter analysis of the regret bound to close the gap between our upper bound and the lower bound of Lai and Robbins (1985), analysis of TS for delayed and batched feedbacks, contextual bandits, prior mismatch and posterior reshaping discussed in Chapelle and Li (2011).

References

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

O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In *NIPS*, 2011.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory (COLT)*, 2011.

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

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.

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.

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.

E. Kaufmann, O. Cappé, and A. Garivier. On bayesian upper confidence bounds for bandit problems. In *Fifteenth International Conference on Artificial Intelligence and Statistics (AISTAT)*, 2012.

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

O.-A. Maillard, R. Munos, and G. Stoltz. Finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Conference on Learning Theory (COLT)*, 2011.

B. C. May and D. S. Leslie. Simulation studies in optimistic bayesian sampling in contextual-bandit problems. Technical Report 11:02, Statistics Group, Department of Mathematics, University of Bristol, 2011.

B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic bayesian sampling in contextual-bandit problems. Technical Report 11:01, Statistics Group, Department of Mathematics, University of Bristol, 2011.

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

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.

Appendix 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,

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

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.

Appendix 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 (Jogdeo and Samuels (1968)) *Median of the binomial distribution Binomial*(n,p) *is either* $\lfloor np \rfloor$ *or* $\lceil np \rceil$.

Fact 3 ((Chernoff–Hoeffding bounds)) *Let $X_1, ..., X_n$ be random variables with common range $[0, 1]$ and such that $\mathbb{E}[X_t \mid X_1, ..., X_{t-1}] = \mu$. 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 9 *For all $n, p \in [0, 1], \delta \geq 0$,*

$$F^B_{n,p}(np - n\delta) \leq e^{-2n\delta^2}, \;\; 1 - F^B_{n,p}(np + n\delta) \leq e^{-2n\delta^2}, \tag{8}$$

$$1 - F^B_{n+1,p}(np + n\delta) \leq \frac{e^{4\delta}}{e^{2n\delta^2}}. \tag{9}$$

Proof The first result is a simple application of Chernoff–Hoeffding bounds from Fact 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,

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

■

Appendix C. Proofs of Lemmas

C.1. Proof of Lemma 5

Proof In this lemma, we lower bound the probability of $E_2(t)$ by $1 - \frac{2}{T^2}$. Recall that event $E_2(t)$ holds if the following is true:

$$\{\theta_2(t) \leq \mu_2 + \frac{\Delta}{2}\} \text{ or } \{k_2(t) < L\}.$$

Also define $A(t)$ as the event

$$A(t) : \frac{S_2(t)}{k_2(t)} \leq \mu_2 + \frac{\Delta}{4},$$

where $S_2(t), k_2(t)$ denote the number of successes and number of plays respectively of the second arm until time $t - 1$. We will upper bound the probability of $\Pr(\overline{E_2(t)}) = 1 - \Pr(E_2(t))$ as:

$$\begin{aligned} \Pr(\overline{E_2(t)}) &= \Pr(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}, k_2(t) \geq L) \\ &\leq \Pr(\overline{A(t)}, k_2(t) \geq L) + \Pr(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}, k_2(t) \geq L, A(t)). \end{aligned} \tag{10}$$

For clarity of exposition, let us define another random variable $\overline{Z}_{2,M}$, as the average number of successes over the first M plays of the second arm. More precisely, let random variable $Z_{2,m}$ denote the output of the m^{th} play of the second arm. Then,

$$\overline{Z}_{2,M} = \frac{1}{M}\sum_{m=1}^{M} Z_{2,m}.$$

Note that by definition, $\overline{Z}_{2,k_2(t)} = \frac{S_2(t)}{k_2(t)}$. Also, $\overline{Z}_{2,M}$ is the average of M iid Bernoulli variables, each with mean μ_2.
Now, for all t,

$$\begin{aligned}
\Pr(\overline{A(t)}, k_2(t) \geq L) &= \textstyle\sum_{\ell=L}^{T} \Pr(\overline{Z}_{2,k_2(t)} \geq \mu_2 + \frac{\Delta}{4}, k_2(t) = \ell) \\
&= \textstyle\sum_{\ell=L}^{T} \Pr(\overline{Z}_{2,\ell} \geq \mu_2 + \frac{\Delta}{4}, k_2(t) = \ell) \\
&\leq \textstyle\sum_{\ell=L}^{T} \Pr(\overline{Z}_{2,\ell} \geq \mu_2 + \frac{\Delta}{4}) \\
&\leq \textstyle\sum_{\ell=L}^{T} e^{-2\ell\Delta^2/16} \\
&\leq \frac{1}{T^2}.
\end{aligned}$$

The second last inequality is by applying Chernoff bounds, since $\overline{Z}_{2,\ell}$ is simply the average of ℓ iid Bernoulli variables each with mean μ_2.

We will derive the bound on second probability term in (10) in a similar manner. It will be useful to define $W(\ell, z)$ as a random variable distributed as Beta$(\ell z+1, \ell-\ell z+1)$. Note that if at time t, the number of plays of second arm is $k_2(t) = \ell$, then $\theta_2(t)$ is distributed as Beta$(\ell\overline{Z}_{2,\ell}+1, \ell-\ell\overline{Z}_{2,\ell}+1)$, i.e. same as $W(\ell, \overline{Z}_{2,\ell})$.

$$\begin{aligned}
\Pr(\theta_2(t) > \mu_2 + \frac{\Delta}{2}, A(t), k_2(t) \geq L) &= \sum_{\ell=L}^{T} \Pr(\theta_2(t) > \mu_2 + \frac{\Delta}{2}, A(t), k_2(t) = \ell) \\
&\leq \sum_{\ell=L}^{T} \Pr(\theta_2(t) > \frac{S_2(t)}{k_2(t)} - \frac{\Delta}{4} + \frac{\Delta}{2}, k_2(t) = \ell) \\
&= \sum_{\ell=L}^{T} \Pr(W(\ell, \overline{Z}_{2,\ell}) > \overline{Z}_{2,\ell} + \frac{\Delta}{4}, k_2(t) = \ell) \\
&\leq \sum_{\ell=L}^{T} \Pr(W(\ell, \overline{Z}_{2,\ell}) > \overline{Z}_{2,\ell} + \frac{\Delta}{4}) \\
\text{(using Fact 1)} \quad &= \sum_{\ell=L}^{T} \mathbb{E}\left[F^{B}_{\ell+1, \overline{Z}_{2,\ell}+\frac{\Delta}{4}}(\ell\overline{Z}_{2,\ell})\right] \\
&\leq \sum_{\ell=L}^{T} \mathbb{E}\left[F^{B}_{\ell, \overline{Z}_{2,\ell}+\frac{\Delta}{4}}(\ell\overline{Z}_{2,\ell})\right] \\
&\leq \sum_{\ell=L}^{T} \exp\{-\frac{2\Delta^2\ell^2/16}{\ell}\} \\
&\leq Te^{-2L\Delta^2/16} = \frac{1}{T^2}.
\end{aligned}$$

The third-last inequality follows from the observation that

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

And, the second-last inequality follows from Chernoff–Hoeffding bounds (refer to Fact 3 and Lemma 9). ■

C.2. Proof of Lemma 6

Proof Using Lemma 4, the expected value of $X(j, s(j), y)$ for any given $s(j)$,

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

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

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,

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

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

$$\mathbb{E}\left[\mathbb{E}\left[X(j, s(j), y) \,|\, s(j)\right]\right] = \mathbb{E}\left[\frac{1}{F^B_{j+1,y}(s(j))} - 1\right] = \sum_{s=0}^{j} \frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} - 1, \tag{11}$$

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 $\lceil np \rceil$ (see Jogdeo and Samuels (1968)). Therefore,

$$\sum_{s=\lceil y(j+1) \rceil}^{j} \frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} \leq 2. \tag{12}$$

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

$$\begin{aligned}
\sum_{s=0}^{\lfloor yj \rfloor} \frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} &\leq \sum_{s=0}^{\lfloor yj \rfloor} \frac{1}{(1-y)} \frac{f^B_{j,\mu_1}(s)}{f^B_{j,y}(s)} \\
&= \sum_{s=0}^{\lfloor yj \rfloor} \frac{1}{(1-y)} \frac{\mu_1^s(1-\mu_1)^{j-s}}{y^s(1-y)^{j-s}} \\
&= \sum_{s=0}^{\lfloor yj \rfloor} \frac{1}{(1-y)} R^s \frac{(1-\mu_1)^j}{(1-y)^j} \\
&= \frac{1}{(1-y)} \left(\frac{R^{\lfloor yj \rfloor + 1} - 1}{R-1} \right) \frac{(1-\mu_1)^j}{(1-y)^j} \\
&\leq \frac{1}{(1-y)} \frac{R}{R-1} \frac{\mu_1^{yj}(1-\mu_1)^{(j-yj)}}{y^{yj}(1-y)^{j-yj}} \\
&= \frac{\mu_1}{\mu_1 - y} e^{-Dj} = \frac{\mu_1}{\Delta'} e^{-Dj}. \qquad (13)
\end{aligned}$$

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 $\lceil yj \rceil \leq yj + y$. For $s = \lceil yj \rceil$,

$$\begin{aligned}
\frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} &\leq \frac{1}{(1-y)F^B_{j,y}(s)} \\
&\leq \frac{2}{1-y}. \qquad (14)
\end{aligned}$$

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

$$\begin{aligned}
\frac{f^B_{j,\mu_1}(s)}{F^B_{j+1,y}(s)} &\leq \frac{1}{(1-y)} \frac{f^B_{j,\mu_1}(s)}{F^B_{j,y}(s)} \\
&\leq \frac{1}{(1-y)} \frac{f^B_{j,\mu_1}(s)}{f^B_{j,y}(s)} \\
&\leq \frac{1}{(1-y)} R^s \left(\frac{1-\mu_1}{1-y} \right)^j \\
&\leq \frac{1}{(1-y)} R^{yj+y} \left(\frac{1-\mu_1}{1-y} \right)^j \qquad \text{(because } s = \lceil yj \rceil \leq yj + y\text{)} \\
&\leq \frac{R^y}{(1-y)} e^{-Dj}. \qquad (15)
\end{aligned}$$

Next, we substitute the bounds from (12)-(15) in Equation (11) to get the result in the lemma. In this substitution, for $s = \lceil yj \rceil$, we use the bound in Equation (14) when $j < \frac{y}{D} \ln R$, and the bound in Equation (15) when $j \geq \frac{y}{D} \ln R$. ■

C.3. Details of Equation (1)

Using Lemma 6 for $y = \mu_2 + \Delta/2$, and $\Delta' = \Delta/2$, we can bound the expected number of plays of the second arm as:

$$\begin{aligned}
\mathbb{E}\left[k_2(T)\right] &= L + \mathbb{E}\left[\sum_{j=j_0}^{T-1} Y_j\right] \\
&\leq L + \sum_{j=0}^{T-1} \mathbb{E}\left[\min\{\mathbb{E}\left[X(j, s(j), \mu_2 + \frac{\Delta}{2}) \middle| s(j)\right],\ T\}\right] + \sum_t \Pr(\overline{E_2(t)}) \cdot T \\
&\leq L + \frac{4\ln T}{\Delta'^2} + \sum_{j=0}^{4(\ln T)/\Delta'^2 - 1} \frac{\mu_1}{\Delta'} e^{-Dj} + \left(\frac{y}{D}\ln R\right)\frac{2}{1-y} + \sum_{j=\frac{y}{D}\ln R}^{4(\ln T)/\Delta'^2 - 1} \frac{R^y e^{-Dj}}{1-y} + \frac{16}{T}\cdot T + 2 \\
&= L + \frac{4\ln T}{\Delta'^2} + \sum_{j=0}^{4(\ln T)/\Delta'^2 - 1} \frac{\mu_1}{\Delta'} e^{-Dj} + \frac{y}{D}\ln R \cdot \frac{2}{(1-y)} + \sum_{j=0}^{4\ln T/\Delta'^2 - \frac{y}{D}\ln R - 1} \frac{1}{1-y} e^{-Dj} + 18 \\
&\leq L + \frac{4\ln T}{\Delta'^2} + \frac{y}{D}\ln R \cdot \frac{2}{\Delta'} + \sum_{j=0}^{T-1} \frac{(\mu_1+1)}{\Delta'} e^{-Dj} + 18 \\
&\overset{(*)}{\leq} L + \frac{4\ln T}{\Delta'^2} + \frac{D+1}{\Delta' D}\cdot\frac{2}{\Delta'} + \frac{2}{\Delta'}\frac{2}{(\min\{D,1\})} + 18 \\
&\overset{(**)}{\leq} L + \frac{4\ln T}{\Delta'^2} + \frac{2}{\Delta'^2} + \frac{1}{\Delta'^4} + \frac{4}{\Delta'^3} + 18 \\
&= L + \frac{16\ln T}{\Delta^2} + \frac{8}{\Delta^2} + \frac{16}{\Delta^4} + \frac{32}{\Delta^3} + 18 \\
&\leq \frac{40\ln T}{\Delta^2} + \frac{48}{\Delta^4} + 18.
\end{aligned}$$

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 $D \geq 2\Delta'^2$.

C.4. Proof of Lemma 7

Proof The proof of this lemma follows on the similar lines as the proof of Lemma 5 in Appendix C.1 for the two arms case. We will prove the second statement, the first statement will follow as a corollary.

To prove the second statement of this lemma, we are required to lower bound the probability of $\Pr(E(t)|s(j) = s)$ for all $t, j, s \leq j$, by $1 - \frac{4(N-1)}{T^2}$, where $s(j)$ denotes the number of successes in first j plays of the first arm. Recall that event $E(t)$ holds if the following is true:

$$\{\forall i \in C(t), \theta_i(t) \in [\mu_i - \frac{\Delta_i}{2}, \mu_i + \frac{\Delta_i}{2}]\}$$

Let us define $E_i^+(t)$ as the event $\{\theta_i(t) \leq \mu_i + \frac{\Delta_i}{2}$ or $i \notin C(t)\}$, and $E_i^-(t)$ as the event $\{\theta_i(t) \geq \mu_i - \frac{\Delta_i}{2}$ or $i \notin C(t)\}$. Then, we can bound $\Pr(\overline{E(t)}|s(j))$ as

$$\Pr(\overline{E(t)}|s(j)) \leq \sum_{i=2}^{N} \Pr(\overline{E_i^+(t)}|s(j)) + \Pr(\overline{E_i^-(t)}|s(j)).$$

Now, observe that

$$\Pr(\overline{E_i^+(t)}|s(j)) = \Pr(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}, k_i(t) \geq L_i|s(j)),$$

where $k_i(t)$ is the number of plays of arm i until time $t-1$.

As in the case of two arms, define $A_i(t)$ as the event

$$A_i(t) : \frac{S_i(t)}{k_i(t)} \leq \mu_i + \frac{\Delta}{4},$$

where $S_i(t), k_i(t)$ denote the number of successes and number of plays respectively of the i^{th} arm until time $t-1$.

We will upper bound the probability of $\Pr(\overline{E_i^+(t)}|s(j))$ for all $t, j, i \neq 1$, using,

$$\begin{aligned} \Pr(\overline{E_i^+(t)}|s(j)) &= \Pr(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}, k_i(t) \geq L_i|s(j)) \\ &\leq \Pr(\overline{A_i(t)}, k_i(t) \geq L_i|s(j)) + \Pr(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}, k_i(t) \geq L_i, A_i(t)|s(j)) \end{aligned} \tag{16}$$

For clarity of exposition, similar to the two arms case, for every $i = 1, \ldots, N$ we define variables $\{Z_{i,m}\}$, and $\overline{Z}_{i,M}$. $Z_{i,m}$ denote the output of the m^{th} play of the i^{th} arm. And,

$$\overline{Z}_{i,M} = \frac{1}{M} \sum_{m=1}^{M} Z_{i,m}$$

Note that for all i, m, $Z_{i,m}$ is Bernoulli variable with mean μ_i, and all $Z_{i,m}, i = 1, \ldots, N, m = 1, \ldots, T$ are independent of each other.

Now, instead of bounding the first term $\Pr(\overline{A_i(t)}, k_i(t) \geq L_i|s(j))$, we prove a bound on $\Pr(\overline{A(t)}, k_2(t) \geq L|Z_{1,1}, \ldots, Z_{1,j})$. Note that the latter bound is stronger, since $s(j)$ is simply $\sum_{m=1}^{j} Z_{1,m}$.

Now, for all t, $i \neq 1$,

$$\begin{aligned}
\Pr(\overline{A_i(t)}, k_i(t) \geq L_i | Z_{1,1}, \ldots, Z_{1,j}) &= \sum_{\ell=L}^{T} \Pr(\overline{Z}_{i,k_i(t)} > \mu_i + \tfrac{\Delta_i}{4}, k_i(t) = \ell | Z_{1,1}, \ldots, Z_{1,j}) \\
&= \sum_{\ell=L}^{T} \Pr(\overline{Z}_{i,\ell} > \mu_i + \tfrac{\Delta_i}{4}, k_i(t) = \ell | Z_{1,1}, \ldots, Z_{1,j}) \\
&\leq \sum_{\ell=L}^{T} \Pr(\overline{Z}_{i,\ell} > \mu_i + \tfrac{\Delta_i}{4} | Z_{1,1}, \ldots, Z_{1,j}) \\
&= \sum_{\ell=L}^{T} \Pr(\overline{Z}_{i,\ell} > \mu_i + \tfrac{\Delta_i}{4}) \\
&\leq \sum_{\ell=L}^{T} e^{-2\ell\Delta_i^2/16} \\
&\leq \tfrac{1}{T^2}
\end{aligned}$$

The third last equality holds because for all i, i', m, m', $Z_{i,m}$ and $Z_{i',m'}$ are independent of each other, which means $\overline{Z}_{i,\ell}$ is independent of $Z_{1,m}$ for all $m = 1, \ldots, j$. The second last inequality is by applying Chernoff bounds, since $\overline{Z}_{i,\ell}$ is simply the average of ℓ iid Bernoulli variables each with mean μ_2.

We will derive the bound on second probability term in (16) in a similar manner. As before, it will be useful to define $W(\ell, z)$ as a random variable distributed as Beta$(\ell z + 1, \ell - \ell z + 1)$. Note that if at time t, the number of plays of arm i is $k_i(t) = \ell$, then $\theta_i(t)$ is distributed as Beta$(\ell\overline{Z}_{i,\ell} + 1, \ell - \ell\overline{Z}_{i,\ell} + 1)$, i.e. same as $W(\ell, \overline{Z}_{i,\ell})$. Now, for the second probability term in (16),

$$\begin{aligned}
&\Pr(\theta_i(t) > \mu_i + \frac{\Delta}{2}, A_i(t), k_i(t) \geq L_i | Z_{1,1}, \ldots, Z_{1,j}) \\
&= \sum_{\ell=L_i}^{T} \Pr(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}, A_i(t), k_i(t) = \ell | Z_{1,1}, \ldots, Z_{1,j}) \\
&\leq \sum_{\ell=L_i}^{T} \Pr(\theta_i(t) > \frac{S_i(t)}{k_i(t)} - \frac{\Delta_i}{4} + \frac{\Delta_i}{2}, k_i(t) = \ell | Z_{1,1}, \ldots, Z_{1,j}) \\
&= \sum_{\ell=L_i}^{T} \Pr(W(\ell, \overline{Z}_{i,\ell}) > \overline{Z}_{i,\ell} + \frac{\Delta_i}{4}, k_i(t) = \ell | Z_{1,1}, \ldots, Z_{1,j}) \\
&\leq \sum_{\ell=L_i}^{T} \Pr(W(\ell, \overline{Z}_{i,\ell}) > \overline{Z}_{i,\ell} + \frac{\Delta_i}{4} | Z_{1,1}, \ldots, Z_{1,j}) \\
&= \sum_{\ell=L_i}^{T} \Pr(W(\ell, \overline{Z}_{i,\ell}) > \overline{Z}_{i,\ell} + \frac{\Delta_i}{4}) \\
\text{(using Fact 1)} \quad &= \sum_{\ell=L_i}^{T} \mathbb{E}\left[F^B_{\ell+1, \overline{Z}_{i,\ell} + \frac{\Delta_i}{4}}(\ell\overline{Z}_{i,\ell}) \right] \\
&\leq \sum_{\ell=L_i}^{T} \mathbb{E}\left[F^B_{\ell, \overline{Z}_{i,\ell} + \frac{\Delta_i}{4}}(\ell\overline{Z}_{i,\ell}) \right] \\
&\leq \sum_{\ell=L_i}^{T} \exp\{-\frac{2\Delta_i^2\ell^2/16}{\ell}\}
\end{aligned}$$

$$\leq \quad Te^{-2L_i\Delta_i^2/16} = \frac{1}{T^2}.$$

Here, we used the observation that for all i, i', m, m', $Z_{i,m}$ and $Z_{i',m'}$ are independent of each other, which means $\overline{Z}_{i,\ell}$ and $W(\ell, \overline{Z}_{i,\ell})$ are independent of $Z_{1,m}$ for all $m = 1, \ldots, j$. The third-last inequality follows from the observation that

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

And, the second-last inequality follows from Chernoff–Hoeffding bounds (refer to Fact 3 and Lemma 9). Substituting above in Equation (16), we get

$$\Pr(\overline{E_i^+(t)}|s(j)) \leq \frac{2}{T^2}$$

Similarly, we can obtain

$$\Pr(\overline{E_i^-(t)}|s(j)) \leq \frac{2}{T^2}$$

Summing over $i = 2, \ldots, N$, we get

$$\Pr(\overline{E(t)}|s(j)) \leq \frac{4(N-1)}{T^2}$$

which implies the second statement of the lemma. The first statement is a simple corollary of this. ■

C.5. Proof of Lemma 8

Proof

$$\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1} V_j^{\ell,a} \,\middle|\, s(j)\right] = \mathbb{E}\left[\sum_{\ell=1}^{T} V_j^{\ell,a} \cdot \mathrm{I}(\gamma_j \geq \ell - 1) \,\middle|\, s(j)\right]$$

Let $\mathcal{F}_{\ell-1}$ denote the history until before the beginning of interval $I_j(\ell)$ (i.e. the values of $\theta_i(t)$ and the outcomes of playing the arms until the time step before the first time step of $I_j(\ell)$). Note that the value of random variable $\mathrm{I}(\gamma_j \geq \ell - 1)$ is completely determined by $\mathcal{F}_{\ell-1}$. Therefore,

$$\begin{aligned}
& \mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1} V_j^{\ell,a} \,\middle|\, s(j)\right] \\
= \quad & \mathbb{E}\left[\sum_{\ell=1}^{T} \mathbb{E}\left[V_j^{\ell,a} \cdot \mathrm{I}(\gamma_j \geq \ell - 1) \;\middle|\, s(j), \mathcal{F}_{\ell-1}\right] \,\middle|\, s(j)\right] \\
= \quad & \mathbb{E}\left[\sum_{\ell=1}^{T} \mathbb{E}\left[V_j^{\ell,a} \;\middle|\, s(j), \mathcal{F}_{\ell-1}\right] \cdot \mathrm{I}(\gamma_j \geq \ell - 1) \,\middle|\, s(j)\right].
\end{aligned}$$

Recall that $V_j^{\ell,a}$ is the number of contiguous steps t for which a is the best arm in saturated set $C(t)$ and iid variables $\theta_1(t)$ have value smaller than $\mu_a + \frac{\Delta_a}{2}$. Observe that given $s(j) = s$ and $\mathcal{F}_{\ell-1}$, $V_j^{\ell,a}$ is the length of an interval which ends when the value of an iid Beta$(s+1, j-s+1)$ distributed variable exceeds $\mu_a + \frac{\Delta_a}{2}$ (i.e., $M(t)$ happens), or if an arm other than a becomes the best saturated arm, or if we reach time T. Therefore, given $s(j), \mathcal{F}_{\ell-1}$, $V_j^{\ell,a}$ is stochastically dominated

by $\min\{X(j,s(j),\mu_a+\frac{\Delta_a}{2}),T\}$, where recall that $X(j,s(j),y)$ was defined as the number of trials until an independent sample from Beta$(s+1,j-s+1)$ distribution exceeds y. That is, for all a,

$$\begin{aligned}\mathbb{E}\left[V_j^{\ell,a}\,\middle|\,s(j),\mathcal{F}_{\ell-1}\right] &\le \mathbb{E}\left[\min\{X(j,s(j),\mu_a+\tfrac{\Delta_a}{2}),T\}\,\middle|\,s(j),\mathcal{F}_{\ell-1}\right]\\ &= \mathbb{E}\left[\min\{X(j,s(j),\mu_a+\tfrac{\Delta_a}{2}),T\}\,\middle|\,s(j)\right].\end{aligned}$$

Substituting, we get,

$$\begin{aligned}&\mathbb{E}\left[\textstyle\sum_{\ell=1}^{\gamma_j+1} V_j^{\ell,a}\,\middle|\,s(j)\right]\\ &\le \mathbb{E}\left[\textstyle\sum_{\ell=1}^{T}\mathbb{E}\left[\min\{X(j,s(j),\mu_a+\frac{\Delta_a}{2}),T\}\,\middle|\,s(j)\right]\cdot\ \mathrm{I}(\gamma_j\ge\ell-1)\ \,\middle|\,s(j)\right]\\ &= \mathbb{E}\left[\min\{X(j,s(j),\mu_a+\tfrac{\Delta_a}{2}),T\}\,\middle|\,s(j)\right]\cdot\mathbb{E}\left[\textstyle\sum_{\ell=1}^{T}\mathrm{I}(\gamma_j\ge\ell-1)\,\middle|\,s(j)\right]\\ &= \mathbb{E}\left[\min\{X(j,s(j),\mu_a+\tfrac{\Delta_a}{2}),T\}\,\middle|\,s(j)\right]\cdot\mathbb{E}\left[\gamma_j+1\,|\,s(j)\right].\end{aligned}$$

This immediately implies,

$$\mathbb{E}\left[\textstyle\sum_{a=2}^{N}\Delta_a\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}V_j^{\ell,a}\,\middle|\,s(j)\right]\right]\le\mathbb{E}\left[\sum_{a=2}^{N}\Delta_a\mathbb{E}\left[\min\{X(j,s(j),\mu_a+\tfrac{\Delta_a}{2}),T\}\,\middle|\,s(j)\right]\cdot\mathbb{E}\left[\gamma_j+1\,|\,s(j)\right]\right]$$

■

Appendix D. Proof of Theorem 2: details

We continue the proof from the main body of the paper.

By (6), regret due to playing saturated arms is bounded by

$$\textstyle\sum_{j=0}^{T-1}\mathbb{E}\left[\mathcal{R}^s(I_j)\right]\ \le\ \sum_{j=0}^{T-1}\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_{a=2}^{N}3\Delta_a V_j^{\ell,a}\right]+2\mathbb{E}\left[\sum_{t\in I_j}I(\overline{E(t)})\right]. \qquad (17)$$

Using Lemma 8, the regret contributed by the first term in (17) is bounded by

$$3\textstyle\sum_{j=0}^{T-1}\mathbb{E}\left[\mathbb{E}\left[\gamma_j\,|\,s(j)\right]\sum_a\Delta_a\mathbb{E}\left[\min\{X(j,s(j),y_a),T\}\,|\,s(j)\right]\right]+\sum_{j=0}^{T-1}\mathbb{E}\left[\mathbb{E}\left[\min\{X(j,s(j),y_a),T\}\,|\,s(j)\right]\right].$$

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 $\mu_i+\frac{\Delta_i}{2}$ of all saturated arms $i\in C(t)$, and yet the first arm was not played. The only reasons the first arm would not be played at a time t despite of $\theta_1(t)>\max_{i\in C(t)}\mu_i+\frac{\Delta_i}{2}$ are that either $E(t)$ was violated, i.e. some saturated arm whose $\theta_i(t)$ was not close to its mean was played instead; or some unsaturated arm u with highest $\theta_u(t)$ was played. Therefore, the random variables γ_j satisfy

$$\gamma_j\le\textstyle\sum_{t\in I_j}I(\text{an unsaturated arm is played at time }t)+\sum_{t\in I_j}I(\overline{E(t)}).$$

Using Lemma 7, and the fact that an unsaturated arm u can be played at most L_u times before it becomes saturated, we obtain that

$$\begin{aligned}\textstyle\sum_{j=0}^{T-1}\mathbb{E}[\gamma_j|s(j)] &\le \textstyle\mathbb{E}[\sum_{t=1}^{T}I(\text{an unsaturated arm is played at time }t)|s(j)]+\sum_{j=0}^{T-1}\mathbb{E}[\sum_{t\in I_j}I(\overline{E(t)})|s(j)]\\ &\le \textstyle\sum_u L_u+\sum_{j=0}^{T-1}\sum_{t=1}^{T}\Pr(\overline{E(t)}|s(j))\\ &\le \textstyle\sum_u L_u+4(N-1). \qquad (18)\end{aligned}$$

Note that $\sum_{j=0}^{T-1} \mathbb{E}[\gamma_j | s(j)]$ is a r.v. (because of random $s(j)$), and the above bound applies for all instantiations of this r.v.

Let $y_a = \mu_a + \frac{\Delta_a}{2}$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=0}^{T-1} \mathbb{E}\left[\gamma_j \,|\, s(j)\right] \sum_a \Delta_a \mathbb{E}\left[X(j, s(j), y_a) \,|\, s(j)\right]\right] \\
&\leq \mathbb{E}\left[\left(\sum_{j=0}^{T-1} \mathbb{E}\left[\gamma_j \,|\, s(j)\right]\right) \left(\max_j \sum_a \Delta_a \mathbb{E}\left[X(j, s(j), y_a) \,|\, s(j)\right]\right)\right] \\
&\leq \left(\sum_u L_u + 4(N-1)\right) \sum_a \Delta_a \mathbb{E}\left[\max_j \mathbb{E}\left[X(j, s(j), y_a) \,|\, s(j)\right]\right] \\
&\leq \left(\sum_u L_u + 4(N-1)\right) \sum_a \Delta_a \mathbb{E}\left[\frac{\Delta_a}{F_{j_a^*+1, y_a}(s(j_a^*))} \cdot I(s(j_a^*) \leq \lfloor y_a j_a^* \rfloor) + \frac{\Delta_a}{F_{j_a^*+1, y_a}(s(j_a^*))} \cdot I(s(j_a^*) \geq \lceil y_a j_a^* \rceil)\right],
\end{aligned} \tag{19}$$

where

$$j_a^* = \arg \max_{j \in \{0, \ldots, T-1\}} \mathbb{E}\left[X(j, s(j), y_a) \,|\, s(j)\right] = \arg \max_{j \in \{0, \ldots, T-1\}} \frac{1}{F_{j+1, y_a}(s(j))}.$$

Note that j_a^* is a random variable, which is completely determined by the instantiation of random sequence $s(1), s(2), \ldots$.

For the first term in Equation (19),

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{F_{j_a^*+1, y_a}(s(j_a^*))} \cdot I(s(j_a^*) \leq \lfloor y_a j_a^* \rfloor)\right] &\leq \sum_j \mathbb{E}\left[\frac{1}{F_{j+1, y_a}(s(j))} \cdot I(s(j) \leq \lfloor y_a j \rfloor)\right] \\
&= \sum_j \sum_{s=0}^{\lfloor y_a j \rfloor} \frac{f_{j, \mu_1}(s)}{F_{j+1, y_a}(s)} \leq \sum_j \frac{\mu_1}{\Delta_a'} e^{-D_a j} \leq \frac{16}{\Delta_a^3}
\end{aligned} \tag{20}$$

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 (13) in the proof of Lemma 6 in Appendix C.2, 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).

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

And, for the second term, using 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),

$$\mathbb{E}\left[\frac{1}{F_{j_a^*+1, y_a}(s(j_a^*))} \cdot I(s(j_a^*) \geq \lceil y_a j_a^* \rceil)\right] \leq \frac{2}{1 - y_a} \leq \frac{4}{\Delta_a}. \tag{21}$$

Substituting the bound from Equation (20) and (21) in Equation (19),

$$\sum_{j=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[\gamma_j | s(j)\right] \sum_a 3\Delta_a \mathbb{E}\left[X(j, s(j), y_a) | s(j)\right]\right] \leq \left(\sum_u L_u + 4(N-1)\right) \sum_a \left(\frac{48}{\Delta_a^2} + 12\right). \tag{22}$$

Also, using Lemma 6 while substituting y with $y_a = \mu_a + \frac{\Delta_a}{2}$ and Δ' with $\mu_1 - y_a = \frac{\Delta_a}{2}$,

$$\begin{aligned}
&\sum_{j=0}^{T-1}\sum_{a=2}^{N}(3\Delta_a)\mathbb{E}\left[\mathbb{E}\left[\min\{X(j,s(j),\mu_a+\frac{\Delta_a}{2}),T\}\,\middle|\,s(j)\right]\right] \\
&\qquad\leq \sum_a(3\Delta_a)\sum_{j=0}^{\frac{16(\ln T)}{{\Delta_a}^2}-1}\left(1+\frac{2}{1-y_a}\right)+\sum_{j\geq\frac{16(\ln T)}{{\Delta_a}^2}}^{T}(3\Delta_a)\frac{16}{T} \\
&\qquad\leq \sum_a\frac{48\ln T}{\Delta_a}+\frac{192}{\Delta_a^2}+48\Delta_a. \qquad (23)
\end{aligned}$$

Substituting bounds from (22) and (23) in the first term of Equation (17),

$$\begin{aligned}
&\sum_{j=0}^{T-1}\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_a V_j^{\ell,a}3\Delta_a\right] \\
&\leq (\sum_u L_u+4(N-1))\sum_a(\frac{48}{\Delta_a^2}+12)+\sum_a(\frac{48\ln T}{\Delta_a}+\frac{192}{\Delta_a^2}+48\Delta_a) \\
&\leq 1152(\ln T)(\sum_i\frac{1}{\Delta_i^2})^2+288(\ln T)\sum_i\frac{1}{\Delta_i^2}+48(\ln T)\sum_a\frac{1}{\Delta_a}+192N\sum_a\frac{1}{\Delta_a^2}+96(N-1).
\end{aligned}$$

Now, using the result that $\Pr(\overline{E(t)}) \leq 4(N-1)/T^2$ (by Lemma 7) with Equation (17), we can bound the total regret due to playing saturated arms as

$$\begin{aligned}
\mathbb{E}[\mathcal{R}^s(T)] &= \sum_j\mathbb{E}[\mathcal{R}^s(I_j)] \\
&= \sum_j\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_a V_j^{\ell,a}3\Delta_a\right]+2T\cdot\sum_t\Pr(\overline{E(t)}) \\
&\leq 1152(\ln T)(\sum_i\frac{1}{\Delta_i^2})^2+288(\ln T)\sum_i\frac{1}{\Delta_i^2} \\
&\qquad +48(\ln T)\sum_a\frac{1}{\Delta_a}+192N\sum_a\frac{1}{\Delta_a^2}+96(N-1)+8(N-1).
\end{aligned}$$

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

$$\mathbb{E}[\mathcal{R}^u(T)] \leq \sum_{u=2}^{N}L_u\Delta_u = 24(\ln T)\left(\sum_{u=2}^{N}\frac{1}{\Delta_u}\right).$$

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

The proof for the alternate bound in Remark 3 will essentially follow the same lines except that instead of dividing the interval $I_j(\ell)$ into subdivisions $V_j^{\ell,a}$, we will simply bound the regret due to saturated arms by number of plays times Δ_{max}. That is, we will use the bound,

$$\mathbb{E}[\mathcal{R}(I_j)] \leq \mathbb{E}[\sum_{\ell=1}^{\gamma_j+1} |I_j(\ell)| \cdot \Delta_{max}]$$

To bound $\mathbb{E}[\sum_{\ell=1}^{\gamma_j+1} |I_j(\ell)|]$, we follow the proof for bounding $\mathbb{E}[\sum_{\ell=1}^{\gamma_j+1} V_j^{\ell,\bar{a}}]$ for $\bar{a} = \arg\max_{i \neq 1} \mu_i$, i.e., replacing μ_a with $\mu_{\bar{a}} = \max_{i \neq 1} \mu_1$, and Δ_a with Δ_{min}. In a manner similar to Lemma 8, we can obtain

$$\mathbb{E}[\sum_{\ell=1}^{\gamma_j+1} |I_j(\ell)|] \leq \mathbb{E}[(\gamma_j + 1) \min\{X(j, s(j), \mu_M + \frac{\Delta_{min}}{2}), T\}] + \mathbb{E}[\sum_{t \in I_j} T \cdot I(\overline{E(t)})]$$

And, consequently, using Equation (19), and Equation (20)–(23), and Lemma 7, we can obtain

$$\sum_j \mathbb{E}[\sum_{\ell=1}^{\gamma_j+1} |I_j(\ell)|] \leq O((\sum_u L_u) \frac{1}{\Delta_{min}^3}) = O(\frac{1}{\Delta_{min}^3} \left(\sum_{a=2}^{N} \frac{1}{\Delta_a^2} \right) \ln T),$$

giving a regret bound of $O(\frac{\Delta_{max}}{\Delta_{min}^3} \left(\sum_{a=2}^{N} \frac{1}{\Delta_a^2} \right) \ln T)$.