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

We show that the Thompson Sampling algorithm achieves logarithmic expected regret for the Bernoulli multi-armed bandit problem. More precisely, for the two-armed bandit problem, the expected regret in time T is $O(\frac{\ln T}{\Delta} + \frac{1}{\Delta^3})$. And, for the 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.

1 Introduction

Multi-armed bandit problems model the exploration/exploitation dilemma inherent in sequential decision problems. Many versions and generalizations of this problem have been studied; in this short paper we will confine ourselves to Bernoulli bandits for the sake of space and exposition. Among many algorithms available for bandit problems, some popular ones include Upper Confidence Bound (UCB) family of algorithms [6, 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 bandit problems, Thompson [10] proposed a natural 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's a member of the family of *randomized probability matching* algorithms.

TS has recently attracted considerable attention. Several studies [9, 2, 8] have empirically demonstrated the efficacy of TS. Scott [9] provides a detailed discussion of probability matching techniques in various 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 [6] for the multi-armed bandit problem; and in applications such as 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 than the other existing methods.

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 could be its lack of strong theoretical analysis. Existing theoretical analyses [4, 7] 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 in time T that is close to the lower bound of [6].

2 Thompson Sampling

In the N-armed Bernoulli bandit problem, each arm yields a reward of 0 or 1, with $\mu_1, \ldots, \mu_N$ being the respective probabilities of getting a reward of 1. W.l.o.g, assume that the first arm is the optimal arm, i.e., $\mu_1 = \max_i \mu_i$; we will refer to the rest of the arms as suboptimal arms. Set $\Delta_i := \mu_1 - \mu_i$.

At every time step t, one of the arms is selected to be played. Let $S_i(t), F_i(t)$ denote the number of successes (reward = 1) and number of failures (reward = 0), respectively, observed for arm i until time t, and let $k_i(t) = S_i(t) + F_i(t)$ denote the number of times arm i has been played so far.

Then, Thompson Sampling algorithm works as follows: At time step t, for each arm i, generate $\theta_i(t)$ from beta distribution with parameters $(S_i(t)+1, F_i(t)+1)$. Play arm $i^*(t)$, where

$$i^*(t) = \arg\max_i \theta_i(t).$$

The expected regret of this algorithm in time T is given by

$$\mathbb{E}[R(T)] = \mathbb{E}[\textstyle\sum_{t=1}^{T} \mu_1 - \mu_{i^*(t)}] = \sum_{i=2}^{N} \Delta_i \mathbb{E}[k_i(T)].$$

3 Our result

In this article, we bound the *finite time* expected regret of Thompson Sampling.

Theorem 1. *For two-armed Bernoulli bandit problem, Thompson Sampling algorithm has expected regret of*

$$\mathbb{E}[R(T)] \leq \tfrac{20}{\Delta} \ln T + \tfrac{64}{\Delta^3} + 2\Delta$$

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

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

$$\mathbb{E}[R(T)] = \left[125(\textstyle\sum_{i=2}^{N} \frac{1}{\Delta_i^2})^2\right] \ln T$$

in time T.

We remark that we have not attempted to optimize the constants in above theorems. Let us contrast our bound with the previous work. Lai and Robbins [6] proved the following lower bound on regret of any bandit algorithm:

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

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:

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

For many settings of the parameters, this bound is not far from the optimal bound of Lai and Robbins. Our bound closely matches the bound of Auer et al. for the two-arms setting. For the N-arms setting, while our bound is slightly inferior to Auer et al. due to the appearance of Δ_i^4 in the denominator instead of Δ_i, it demonstrates that Thompson algorithm achieves logarithmic regret even in the case of more than two arms.

Proof Techniques Here we give an informal description of the high-level ideas involved in our analysis.

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 in the time steps *after* the second arm has been played $L = 4(\ln T)/\Delta^2$ times. The expected regret before this event is bounded by $4(\ln T)/\Delta$ because only the plays of second arm produce an expected regret of Δ, the 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 suboptimal arm is very close to its actual expected reward μ_2, and its beta distribution is tightly concentrated around μ_2. This means that 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 $\Pr[\theta_1(t) \geq \mu_2]$. 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 step in which the first arm is played). This is modeled by a geometric random variable $X(j, s(j), \mu_2)$ with parameter $\Pr[\theta_1 \geq \mu_2]$, where θ_1 has distribution $\text{Beta}(s(j)+1, j-s(j)+1)$, and thus $\mathbb{E}[X(j, s(j), \mu_2)] = 1/\Pr[\theta_1 \geq \mu_2] - 1$. To bound the overall expected number of steps between j and $j+1$ play of 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 expected number of steps between consecutive plays of first arm bounds the expected number of plays of second arm to yield a good bound on the regret for the two-arms setting.

Next, we consider the case of more than two arms ($N > 2$), i.e., the case of multiple suboptimal arms. A natural approach to extend the analysis of single suboptimal arm case to multiple suboptimal arms would be to bound the number of plays of each suboptimal arm by the number of times it exceeds the first arm. This approach is taken in Auer et al. [1] for analyzing UCB algorithm. A main difficulty in decomposing the multiple suboptimal arms case in this manner is that we could be wrongly attributing all the plays other than the plays of suboptimal arm being considered, to be the plays of the first arm, thus over-counting the previous plays of the first arm. Auer et al. overcome this difficulty by showing that for *all possible number of previous plays* of the first arm, the probability of playing a suboptimal arm that has already been played large enough times at time t, is very small (inverse polynomial in t). However, for Thompson Sampling, if the number of previous plays of the first arm is small at time t, then the probability of a suboptimal arm i having a greater $\theta_i(t)$ than $\theta_1(t)$ can be as large as a constant, even if this suboptimal arm has already been played $O(\ln T/\Delta^2)$ times: for example, 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 $\theta_2(t) \approx \mu_2$. Thus, the (distribution of the) number of previous plays of first arm needs to be carefully accounted for, which requires a more involved analysis of the multiple suboptimal arms case.

The main ideas in our analysis of this case are as follows. 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 ($L_a = 4 \ln T/\Delta_a^2$) 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 the j^{th} play, the $(j+1)^{th}$ play of the first arm will occur at the earliest time t such that $\theta_1(t) \geq \theta_i(t), \forall i$. The number of steps before $\theta_1(t)$ is greater than $\theta_a(t)$ of a saturated arm a can be analyzed using geometric random variable with parameter close to $\Pr(\theta_1 \geq \mu_a)$ as earlier. 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 i with a greater $\theta_i(t)$. Call this event an "interruption" by unsaturated arms. We show that given that 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 $X(j, s(j), \mu_a)$ defined earlier, and number of interruptions by the unsaturated arms. Now, the total number of interruptions by unsaturated arms is bounded by $\sum_i L_i$. The actual number of interruptions in an interval 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 $\sum_i L_i$ interruptions.

In this short article, we only present the proof for the two-arms setting, i.e. when $N = 2$.

4 Regret bound for the two-armed bandit problem

In this section, we present a proof of our result for two arms, while omitting many technical details due to space considerations.

Let random variable j_0 denote the number of plays of the first arm until $4 \ln T/\Delta^2$ 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 first arm, and let $s(j)$ be the number of successes in j plays. Then, the expected number of plays of the second arm in time T is bounded by

$$\mathbb{E}[k_2(T)] \leq \frac{4 \ln T}{\Delta^2} + \mathbb{E}[\sum_{j=j_0}^{T} Y_j]$$

Define $X(j, s, y)$ as a geometric random variable denoting the number of steps *before* a $\text{Beta}(s+1, j-s+1)$ distributed random variable exceeds a threshold y for the first time. The following lemma provides a handle on the expectation of X.

Lemma 1. *For all 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 cdf of the Binomial distribution with parameters (n,p).

Proof. By well-known properties of geometric random variables and the definition of X we have, $\mathbb{E}[X(j,s,y)] = \frac{1}{1-F^{beta}_{s+1,j-s+1}(y)} - 1$, where $F^{beta}_{\alpha,\beta}$ denotes cdf of the beta distribution with parameters (α,β). (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 the following known relationship between cdf of Binomial and beta distributions:

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

for all integers α,β. □

Define $E(T)$ to be the event that $\theta_2(t) \le \mu_2 + \frac{\Delta}{2}$ for all time $t \in [1,T]$ such that the second arm has already been played at least $4(\ln T)/\Delta^2$ times before time t. The following can be proven using the above relation between the beta and Binomial distributions, and by application of the Chernoff-Hoeffding bounds. We omit the proof.

Lemma 2. $\Pr(E(T))) \ge 1 - \frac{2}{T}$.

Recall that Y_j is defined as the number of steps before $\theta_1(t) \ge \theta_2(t)$ for the first time after the j^{th} play of the first arm. Now, under event $E(T)$, $\theta_2(t) \le \mu_2 + \Delta/2$, for all time t after $4(\ln T)/\Delta^2$ plays of the second arm. Therefore, given that event $E(T)$ holds, and that the number of successes in j trials of first arm is $s(j)$, Y_j for $j \ge j_0$ is dominated by the geometric random variable $X(j,s(j),\mu_2+\frac{\Delta}{2})$. If the event $E(T)$ does not hold, we bound $\sum_j Y_j$ by T.

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

Lemma 3. *Consider any $y < \mu_1$, and let $\Delta' = \mu_1 - y$. Also, let $R = \frac{\mu_1(1-y)}{y(1-\mu_1)}$, 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,*

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

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

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

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

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

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 $s \le \mu_1 j$ with binomial probability $f^B_{j,\mu_1}(s)$. For such s, we use the lower bound $F^B_{j+1,y}(s) \ge (1-y)F^B_{j,y}(s) \ge (1-y)f^B_{j,y}(s)$, and then bound the ratio $f^B_{j,y}(s)/f^B_{j,\mu_1}(s)$ in terms of Δ', R and KL-divergence D. For $s(j) = s \ge \lceil \mu_1 j \rceil$, we use the observation that since $\lceil \mu_1 j \rceil$ is greater than or equal to the median of Binomial(j,μ_1) [5], we have $F^B_{j,y}(s) \ge 1/2$. After some algebraic manipulations, we get the result of the lemma. □

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:

$$\begin{aligned}
\mathbb{E}[k_2(T)] = \frac{4\ln T}{\Delta^2} + \mathbb{E}[\sum_{j=j_0}^{T} Y_j] &\leq \frac{4\ln T}{\Delta^2} + \sum_{j=1}^{T} \mathbb{E}[\ \mathbb{E}[X(j, s(j), \mu_2 + \frac{\Delta}{2}) \mid s(j)]\ ] + 2 \\
&\leq \frac{4\ln T}{\Delta^2} + \frac{4\ln T}{\Delta'^2} + \frac{2}{(1-y)} \cdot \frac{y}{D} \ln \frac{R}{2} + \frac{\mu_1 + 1}{\Delta'} \cdot \frac{1}{(1 - e^{-D})} + 2 \\
&\leq \frac{20\ln T}{\Delta^2} + \frac{64}{\Delta^4} + 2.
\end{aligned}$$

This gives a regret bound of $(\frac{20\ln T}{\Delta} + \frac{64}{\Delta^3} + 2\Delta)$ in case of $N = 2$.

Conclusion In this short article, we demonstrated that theoretical guarantees close to other state of the art methods, like UCB, can be obtained for Thompson Sampling. Our technique allows various extensions such as analysis of Thompson Sampling for bandits more general than Bernoulli bandits, delayed and batched feedbacks, prior mismatch and posterior reshaping; these extensions will be treated in future work.

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*, 2012.

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

[4] 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.

[5] K. Jogdeo and S. M. Samuels. Monotone convergence of binomial probabilities and a generalization of ramanujans equation. *The Annals of Mathematical Statistics*, (4):1191–1195, 1968.

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

[7] 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.

[8] 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.

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

[10] 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.