Thompson Sampling Paper Summarization

Jiapeng Wu, 260727743

January 30, 2019

1 Introduction

Thompson sampling is a randomized algorithm based on Bayesian ideas for multi-armed bandit algorithm. In this paper, the authors provide a regret analysis for Thompson Sampling that proves both the optimal problem-dependent bound of $(1+\epsilon)\sum_i \frac{\ln T}{\Delta_i} + O(\frac{N}{\epsilon^2})$and the first near-optimal problem-independent bound of $O(\sqrt{NTlnT})$ on the expected regret of this algorithm.

Thompson Sampling The basic idea of Thompson Sampling(TS) 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.
The Bernoulli bandit problem is a special case when the rewards are either 0 or 1. The algorithm for Bernoulli bandits maintains Bayesian priors on the Bernoulli means μ_i's. Beta distribution is useful for Bernoulli rewards because if the prior is a $Beta(\alpha, \beta)$ 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.

2 Summary of Proof

Without the loss of generosity, we assume that the first arm is the unique optimal arm.

Theorm 1(Problem-dependent bound) For the N-armed stochastic bandit problem, Thompson Sampling algorithm has expected regret

$$E[R(T)] \leq (1+\epsilon)\sum_i \frac{\ln T}{\Delta_i} + O(\frac{N}{\epsilon^2}) \tag{1}$$

in T is time and $d(u_i, u_1) = u_i log\frac{u_i}{u_1} + (1-u) log\frac{1-u_i}{1-u_1}$

Theorem 1(Problem-independent bound) For the N-armed stochastic bandit problem, Thompson Sampling algorithm has expected regret

$$E[R(T)] \leq O(\sqrt{NT\ln T}) \tag{2}$$

Notations

1. $k_i(t)$ is the number of plays of arm i until time $t-1$. $S_i(t)$ is the number of successes among the plays of arm i until time $t-1$. $i(t)$ denotes the arm played at time t.

2. For each arm i, x_i and y_i are two thresholds such that $\mu_i < x_i < y_i < \mu_1$. Define $L_i(T) = \frac{lnT}{d(x_i,y_i)}$ and $\hat{\mu}_i(t) = S_i(t)/k_i(t)$. Define $E_i^{\mu}(t)$ as the event that $\hat{\mu}_i(t) \leq x_i$. Define $E_i^{\theta}(t)$ as the event that $\theta_i(t) \leq y_i$, where $\theta_i(t)$ is the probability of choosing arm i at time t.

3. Define $F_{t-1} = \{i(w), r_{i(w)}, i = 1, ..., N, w = 1, ..., t-1\}$ where $r_i(t)$ denotes the reward observed for arm i at time t. Define $p_{i,t}$ as $p_{i,t} = Pr(\theta_1(t) > y_i | F_{t-1})$.

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

$$Pr(i(t) = i, E_i^u(t), E_i^{\theta}(t) | \mathcal{F}_{t-1}) \leq \frac{(1 - p_{i,t})}{p_{i,t}} Pr(i(t) = 1 | E_i^u(t), E_i^{\theta}(t) | \mathcal{F}_{t-1}). \tag{3}$$

Lemma 2.

$$\sum_{t=1}^{T} Pr(i(t) = i, \overline{E_i^u(t)}) \leq \frac{1}{d(x_i, \mu_i)} + 1 \tag{4}$$

Lemma 3.

$$\sum_{t=1}^{T} Pr(i(t) = i, \overline{E_i^{\theta}(t)}, E_i^{\mu}(t)) \leq L_i(T) + 1 \tag{5}$$

Lemma 4. Let τ_j denotes the time step at which j^{th} trial of first arm happens, then

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

We proof the two theorems using the four lemmas:

Proof.

$$E[k_i(T)] = \sum_{t=1}^{T} Pr(i(t) = i) \tag{6}$$

$$= \sum_{t=1}^{T} Pr(i(t) = i, E_i^u(t), E_i^{\theta}(t)) + \sum_{t=1}^{T} Pr(i(t) = i, E_i^u(t), \overline{E_i^{\theta}(t)}) + \sum_{t=1}^{T} Pr(i(t) = i, \overline{E_i^u(t)}) \tag{7}$$

$$(*) \leq \sum_{t=1}^{T} E[\frac{1 - p_{i,t}}{p_{i,t}} I(i(t) = 1, E_i^u(t), E_i^{\theta}(t))] + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \tag{8}$$

$$\leq E[\frac{1 - p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \sum_{t=\tau_k+1}^{\tau_{k+1}} I(i(t) = 1)] + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \tag{9}$$

$$= E[\frac{1}{p_{i,\tau_k+1}} - 1] + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \tag{10}$$

$$\leq \frac{24}{\Delta_i'^2} + \sum_{j=0}^{T-1} \Theta(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4} - 1}) + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \tag{11}$$

The inequality marked (*) uses the observation that $p_{i,t}$ changes only when the distribution of $\theta_1(t)$ changes. Thus $p_{i,t}$ is the same at all time steps $t \in \{\tau_k + 1, ..., \tau_{k+1}\}$.

For some $0 < \epsilon \leq 1$, $x_i \in (\mu_i, \mu_1)$ such that $d(x_i, \mu_1) = d(\mu_i, \mu_1)/(1+\epsilon)$ and $y_i \in (x_i, \mu_1)$ such that $d(x_i, y_i) = d(x_i, \mu_1)/(1+\epsilon) = d(\mu_i, \mu_1)/(1+\epsilon)^2$. Then we have

$$L_i(T) = \frac{\ln T}{d(x_i, y_i)} = (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)} \tag{12}$$

after some manipulation we get

$$x_i - \mu_i \geq \frac{\epsilon}{(1+\epsilon)} \frac{d(\mu_i, \mu_1)}{\ln(\frac{\mu_1(1-\mu_i)}{\mu_i(1-\mu_1)})} \rightarrow \frac{1}{d(x_i, \mu_i)} \leq \frac{2}{(x_i - \mu_i)^2} = O(\frac{1}{\epsilon^2}) \tag{13}$$

and

$$\sum_{j=0}^{T-1} \Theta(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4} - 1}) \leq \Theta(\frac{1}{\Delta_i'^2} + \frac{1}{\Delta_i'^2 D}) + \frac{1}{\Delta_i'^4} + \frac{1}{\Delta_i'^2} = \Theta(1) \tag{14}$$

Combining, we get

$$E[R(T)] = \sum_i \Delta_i E[k_i(T)] = \sum_i (1+\epsilon)^2 \frac{\ln T}{d(\mu_i, \mu_1)} \Delta_i + O(\frac{N}{\epsilon^2}) \leq \sum_i (1+\epsilon') \frac{\ln T}{d(\mu_i, \mu_1)} \Delta_i + O(\frac{N}{\epsilon'^2}) \tag{15}$$

Where $\epsilon' = 3\epsilon$

To prove Theorem 2, we pick $x_i = \mu_i + \frac{\Delta_i}{3}$, $y_i = \mu_1 - \frac{\Delta_i}{3}$, so that $\Delta'^2 = (\mu_1 - y_i)^2 = \frac{\Delta_i^2}{9}$. Using Pinsker's inequality, $d(x_i, \mu_i) \geq \frac{1}{2}(x_i - \mu_i)^2 = \frac{\Delta_i^2}{18}$, $d(x_i, y_i) \geq \frac{1}{2}(y_i - x_i)^2 \geq \frac{\Delta_i^2}{18}$. Then

$$L_i(T) = \frac{\ln T}{d(x_i, y_i)} \leq \frac{18 \ln T}{\Delta_i^2} \rightarrow \frac{1}{d(x_i, y_i)} \leq \frac{18}{\Delta_i^2}. \tag{16}$$

$$\begin{aligned} \sum_{j=0}^{T-1} \Theta(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4} - 1}) &\leq \sum_{j=0}^{T-1} \Theta(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} + \frac{4}{\Delta_i'^2 j}) \\ &= \Theta(\frac{1}{\Delta_i'^2} + \frac{\ln T}{\Delta_i'^2}) \\ &= \Theta(\frac{\ln T}{\Delta_i^2}) \end{aligned}$$

This gives,

$$E[k_i(T)] = O(\frac{\ln T}{\Delta_i^2}). \tag{17}$$

$$E[R(T)] = \sum_i \Delta_i E[k_i(T)] = O(\sum_{i \neq 1} \frac{\ln T}{\Delta_i}). \tag{18}$$

The total regret on playing arms with $\Delta_i < \sqrt{\frac{N \ln T}{T}}$ is at most $\sqrt{NT \ln T}$. Thus, for all suboptimal i, $\Delta_i \geq \sqrt{\frac{N \ln T}{T}}$ and all the arms with $\Delta_i < \sqrt{\frac{N \ln T}{T}}$ can be considered as an optimal arm. Substituting $\Delta_i = \sqrt{\frac{N \ln T}{T}}$ to the equation 7 above, we can get the problem-independent upper bound of the expected total regret(Theorem 2).

$$E[\mathcal{R}(\mathcal{T})] = O(\sqrt{NT \ln T})$$

□