EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 22 - 2/9/2021

Lecture 7: Frequentist Regret Bound for Thompson Sampling

Lecturer: Jiantao Jiao *Scribe: Xinlei Pan, Banghua Zhu, Shijie Gu*

The Thompson Sampling algorithm is a heuristic method for dealing with the exploration-exploitation dilemma in multi-armed bandits. The idea is to sample from the posterior of reward distribution and play the optimal action. In this lecture we analyze the frequentist regret bound for Thompson sampling algorithm.

1 Algorithm Description: Bernoulli-Bandit Thompson Sampling

In bandit problems, we usually assume that the reward is bounded in $[0, 1]$. In this lecture we focus on a special case of Bernoulli-Bandit. We start with the description of the Bernoulli-Bandit Thompson Sampling (TS) algorithm.

Bernoulli-Bandit Thompson Sampling. The Bernoulli-Bandit TS(Russo et al. (2017)) problem is a multi-armed bandit problem where there are a total of K actions. Each action, when played, yields either a success (which gives reward 1) or a failure (which gives reward 0) with certain probability, i.e. the reward of each arm follows an independent Bernoulli distribution. The success probabilities for all arms are unknown to the agent, but are fixed over time. The goal is to maximize the cumulative number of successes over T periods, with $T > K$.

We collect necessary notations here.

- $N_i(t)$: denotes the number of pulls of arm i up to time $t-1$;
- $S_i(t)$: denotes the number of successful pulls of arm i up to time $t-1$;
- $F_i(t)$: denotes the number of failed pulls of arm i up to time $t-1$;
- $i(t)$: denotes the arm played at time t;
- $r_i(t)$: denotes the reward of arm i at time t;

So we always have $N_i(t) = S_i(t) + F_i(t)$. Note that in previous lectures we also defined $n_i(t)$, which is the number of pulls of arm i till time t. We recall the Bernoulli-Bandit Thompson sampling algorithm as below(Agrawal and Goyal (2013)):

Algorithm 1: Bernouli-Bandit Thompson Sampling (TS)

```
Initialization: for each arm $i \in [k]$, set $S_i = 0$, $F_i = 0$;
for $t = 1, 2, \cdots, T$ do
    for each arm $i$ do
        sample $\theta_i(t) \sim Beta(S_i + 1, F_i + 1)$
    end
    Play arm $i(t) := \arg\max_j \theta_j(t)$;
    Observe reward $r_i(t)$;
    if $r_i(t) = 1$ then
        $S_i(t) = S_i(t) + 1$
    else
        $F_i(t) = F_i(t) + 1$
    end
end
```

The reason to use Beta distribution for Bernoulli rewards is that the beta distribution is a conjugate prior for the Bernoulli distribution: if the prior is a Beta(α, β) distribution, then after observing a Bernoulli trial, the posterior distribution is Beta($\alpha+1, \beta$) if the trial is a success or Beta($\alpha, \beta+1$) if the trial is a failure. The reason that we have 1 added to both parameters of Beta distribution is that when $S_i(t) = F_i(t) = 0$, the distribution $Beta(1,1)$ is uniform. It's natural to have a uniform prior.

2 Frequentist Regret Analysis

Now we show the main theorem for the frequentist regret bound of Bernoulli-Bandit TS algorithm:

Theorem 1. *(Agrawal and Goyal (2012)) For the Bernoulli-Bandit Thompsong Sampling algorithm described above, we have*

$$\mathbb{E}_V[R(T)] \lesssim \sqrt{KT\log(T)}$$

for any bounded reward family V.

Proof To prove this theorem, we first note that the regret has the following decomposition:

$$R(T) = \sum_{i=1}^{K} N_i(T+1)\Delta_i,$$

where $\Delta_i = \mu^* - \mu_i$, and $\mu^* = \mu_1$ (we assume that arm 1 is the unique best arm). By taking expectation on both sides, we have

$$\mathbb{E}_V[R(T)] = \sum_{i=1}^{K} \mathbb{E}_V(N_i(T+1))\Delta_i.$$

Recall that in the regret analysis for upper confidence bound algorithm, we show that when conditioned on some clean event which happens with high probability, we have $N_i(T+1) \lesssim \frac{\log(T)}{\Delta_i^2}$. In Bernoulli-Bandit TS we aim at bounding the expectation of $N_i(T+1)$, i.e.

$$\mathbb{E}_V[N_i(T+1)] \lesssim \frac{\log(T)}{\Delta_i^2}.$$

To prove the above guarantee, for each arm i we will introduce two intermediate numbers x_i, y_i that satisfies $\mu_i < x_i < y_i < \mu_1$. Later we will use concentration arguments to control the confidence level that is affected by the choice of x_i, y_i. We make the following definitions for the convenience of the proof:

- $\hat{\mu}_i(t) \triangleq \frac{S_i(t)}{N_i(t)}$, initially we have $N_i(t) = 0$, and we define $\hat{\mu}_i(t) = 1$.
- $\theta_i(t)$: the sampled reward at time t for arm i (as above in Algorithm 1)
- $D(x_i||y_i) \triangleq D(Bern(x_i)||Bern(y_i))$: the KL divergence between two Bernoulli distributions with parameter x_i, y_i.
- $L_i(T) \triangleq \frac{\log(T)}{D(x_i||y_i)}$: a standard term that appears naturally in the asymptotic optimality bound. We will see from later proof that this comes from the Chernoff inequality. This is also a generalization of the term $\frac{\log(T)}{\Delta_i}$ in the bandit case.
- $E_i^{\mu}(t) \triangleq \{\hat{\mu}_i(t) \leq x_i\}, E_i^{\theta}(t) \triangleq \{\theta_i(t) \leq y_i\}$: two good events that we expect to happen with high probability since both $\hat{\mu}_i(t)$ and $\theta_i(t)$ concentrates around μ_i, and we know $\mu_i < x_i < y_i$.

- $\mathcal{F}_{t-1} = \{(i(w), r_{i(w)}(w))\}_{1\leq w\leq t-1}$: the filtration for all the actions and rewards we have observed up to time $t-1$. By definition, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \subseteq \mathcal{F}_{T-1}$. Also by definition, for every arm i, the quantities $S_i(t)$, $N_i(t)$, $\hat{\mu}_i(t)$, the distribution of $\theta_i(t)$ and whether $E_i^\mu(t)$ is true or not, are determined by $\mathcal{F}_{t-1}$(Agrawal and Goyal (2013)).
- $P_{i,t} \triangleq P(\theta_1(t) > y_i|\mathcal{F}_{t-1})$: Intuitively, this term also quantifies the probability of good events since $\theta_1(t)$ concentrates around μ_1 and $\mu_1 > y_i$.

With the notations above, we are ready to prove the key lemma for the regret analysis:

Lemma 2. *For $1 \leq t \leq T, i \neq 1$: $P(i(t) = i, E_i^\mu(t), E_i^\theta(t) \mid \mathcal{F}_{t-1}) \leq \frac{1-P_{i,t}}{P_{i,t}} P(i(t) = 1, E_i^\mu(t), E_i^\theta(t) \mid \mathcal{F}_{t-1})$.*

This is lemma 1 in (Agrawal and Goyal (2013)). The term $\frac{1-P_{i,t}}{P_{i,t}}$ is very small given that $P_{i,t}$ is big (close to 1). So the lemma indicates that it is very unlikely to pull the sub-optimal arm since the probability of pulling the sub-optimal arm is significantly smaller than the probability of pulling the optimal arm. We provide the detailed proof of the lemma as below.

Proof Note that $E_i^\mu(t)$ is completely determined by the past observations $\mathcal{F}_{t-1}$. Thus we can always assume this event holds. (Otherwise the probabilities on both sides are 0). Then it suffices to show:

$$P(i(t) = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) \leq \frac{1-P_{i,t}}{P_{i,t}} P(i(t) = 1 \mid E_i^\theta(t), \mathcal{F}_{t-1}).$$

Define the intermediate event that the arm i always sampled to have reward more than all other arms except arm 1: $M_i(t) = \{\theta_i(t) \geq \theta_j(t), \forall j \neq 1\}$, then it suffices to show:

$$\begin{aligned} P(i(t) = 1 \mid E_i^\theta(t), \mathcal{F}_{t-1}) &\geq P_{i,t} \cdot P(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ P(i(t) = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) &\leq (1 - P_{i,t}) \cdot P(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \end{aligned}$$

To see the first inequality, we have

$$\begin{aligned} P(i(t) = 1 \mid E_i^\theta(t), \mathcal{F}_{t-1}) &\geq P(i(t) = 1, M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= P(i(t) = 1|M_i(t), E_i^\theta(t), \mathcal{F}_{t-1}) \cdot P(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \end{aligned}$$

So it suffices to show that $P(i(t) = 1|M_i(t), E_i^\theta(t), \mathcal{F}_{t-1}) \geq P_{i,t}$. Note that under the two events $M_i(t)$ and $E_i^\theta(t)$, the condition $\theta_1 > y_i$ implies that $\theta_1 > y_i > \theta_i \geq \theta_j, \forall j \neq 1$. Thus we must choose action 1 in this case. Therefore,

$$P(i(t) = 1|M_i(t), E_i^\theta(t), \mathcal{F}_{t-1}) \geq P(\theta_1(t) > y_i|M_i(t), E_i^\theta(t), \mathcal{F}_{t-1}) = P(\theta_1(t) > y_i|\mathcal{F}_{t-1}) = P_{i,t}.$$

To see the second inequality, we have

$$\begin{aligned} P(i(t) = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) &\leq P(\theta_1(t) \leq y_i, M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= \underbrace{P(\theta_1(t) \leq y_i \mid E_i^\theta(t), \mathcal{F}_{t-1})}_{=P(\theta_1(t)\leq y_i|\mathcal{F}_{t-1})} P(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= (1 - P_{i,t}) P(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}). \end{aligned}$$

The two inequalities altogether conclude the proof of the lemma. □

To complete the proof of the theorem, we introduce the following additional lemmas without proof:

Lemma 3.

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

where $\overline{A}$ means the complement of an event A.

Lemma 4.

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

where $L_i(T) = \frac{\log(T)}{D(x_i||y_i)}$, $D(x_i||y_i) = D(Bern(x_i)||Bern(y_i))$.

Lemma 5 (Agrawal and Goyal (2013))**.** *Let* τ_j *be the time step we pulled arm* 1 *the* j*th time,* $\Delta_i' = \mu_1 - y_i$, $D_i = D(y_i||\mu_1) = y_i \ln \frac{y_i}{\mu_1} + (1-y_i)\ln\frac{1-y_i}{1-\mu_1}$. *Then*

$$\mathbb{E}\left[\frac{1}{P_{i,\tau_j+1}}\right] \leq \begin{cases} 1 + 3/\Delta_i', & \textit{if } j < 8/\Delta_i' \\ 1 + \Theta(\exp(-\Delta_i'^2 j/2) + \frac{\exp(-D_i j)}{(j+1)\Delta_i'^2} + \frac{1}{\exp(\Delta_i'^2 j/4)-1}), & \textit{otherwise} \end{cases} \tag{1}$$

Now we show how to combine all the lemmas above to prove the main theorem. We hope to upper bound

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

The last two terms are controlled in Lemma 3 and 4. It suffices to control the first term. We have from the main lemma that

$$\begin{aligned} \sum_{t=1}^{T} P(i(t) = 1, E_i^\mu(t), E_i^\theta(t)) &\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-P_{i,t}}{P_{i,t}} 1(i(t) = 1, E_i^\mu(t), E_i^\theta(t))\right] \text{(Lemma 2)} \\ &= \sum_{j=0}^{T-1} \mathbb{E}\left[\frac{1-P_{i,\tau_j+1}}{P_{i,\tau_j+1}} \sum_{t=\tau_j+1}^{\tau_{j+1}} 1(i(t) = i, E_i^\mu(t), E_i^\theta(t))\right] \text{(by definition that } \tau_j\text{)} \\ &\leq \sum_{j=0}^{T-1} \mathbb{E}\left[\frac{1-P_{i,\tau_j+1}}{P_{i,\tau_j+1}} \sum_{t=\tau_j+1}^{\tau_{j+1}} 1(i(t) = i)\right] \text{(drop good events)} \\ &= \sum_{j=0}^{T-1} \mathbb{E}\left[\frac{1}{P_{i,\tau_j+1}} - 1\right] \text{.(by definition of } \tau_j\text{)} \end{aligned}$$

Here in the first inequality we apply Lemma 2; in the second inequality we use the definition that τ_j refers to the jth time arm 1 has been pulled; in the third inequality we drop the good events; and finally in the last equality we use the fact that between time $\tau_j + 1$ and τ_{j+1} the arm 1 is only pulled at time τ_{j+1}.

Now we are ready to prove the $\sqrt{KT\log(T)}$ regret bound, which is deferred to next lecture. □

2.1 Additional Reading Materials

Please refer to Gopalan et al. (2014) and Kveton et al. (2019) for additional readings.

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

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

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

Branislav Kveton, Csaba Szepesvari, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610. PMLR, 2019.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.