STATS 710 – Seq. Dec. Making with mHealth Applications Sep 22, 2016

Lecture 6: Proof of The Upper Bound for Thompson Sampling with Beta Priors and Bernoulli Rewards

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Young Jung*

1 Thompson Sampling with Beta Priors and Bernoulli Rewards

1.1 Algorithm Description And Theorem Statement

Recall from last lecture:

Algorithm 1 Thompson Sampling for Bernoulli Bandits

1: Using beta distribution as prior.
2: For each arm a $\in \mathcal{A}$, $S_a = F_a = 0$.
3: **for** t=1,2, $\cdots$, **do**
4: For each a$\in \mathcal{A}$, sample $\theta_{a,t} \sim Beta(S_a + 1, F_a + 1)$.
5: Play arm $A_t = \text{argmax}_{a\in\mathcal{A}}\ \theta_{a,t}$ and collect reward R_t.
6: If $R_t = 1$, $S_{A_t} = S_{A_t} + 1$. Else if $R_t = 0$, $F_{A_t} = F_{A_t} + 1$.
7: **end for**

This algorithm dates back to 1933 [Tho33], but the first finite time regret analysis appeared only in 2012 [AG12]. Our presentation of TS and its proof will follow a later paper [AG13].

Theorem 1. *Assume Bernoulli reward distributions for the K arms with means $\mu_0, \cdots, \mu_{K-1}$, where $\mu_0 > max_{a\geq 1}\mu_a$. Then for any ε,*

$$R_T(\textit{TS with Beta}, (\mathcal{D}_a)_{a\in\mathcal{A}}) \leq (1+\varepsilon)\sum_{a=1}^{K-1}\frac{\log T}{d(\mu_a,\mu_0)}\Delta_a + O(\frac{K}{\varepsilon^2})$$

where $d(\mu_a, \mu_0) := \mu_a \log\frac{\mu_a}{\mu_0} + (1-\mu_a)\log\frac{1-\mu_a}{1-\mu_0}$.

Note: By Pinsker's inequality, $d(\mu_a, \mu_0) \geq 2(\mu_a - \mu_0)^2 = 2\Delta_a^2$. The upper bound for UCB is worse than that of TS due to this inequality.

1.2 Notation and comments

- $\theta_{a,t}$: draw from arm a's posterior at time t.
- $S_{a,t}$: the number of successes (i.e., ones) of arm a up to and including time t.
- $F_{a,t}$: the number of failures (i.e., zeroes) of arm a up to and including time t.
- Note that $S_{a,t} + F_{a,t} = N_t(a)$, which is the number of times that arm a is selected up to and including time t.
- $\hat{\mu}_{a,t-1} = \frac{\sum_{i=1,A_i=a}^{t-1} R_i^a}{N_{t-1}(a)+1} = \frac{S_{a,t-1}}{N_{t-1}(a)+1}$

- Note that 1 is added to the denominator on purpose to prevent division by zero.

- For each a and any $\varepsilon > 0$, we will introduce thresholds x_a, y_a satisfying following conditions.
  - $\mu_a < x_a < y_a < \mu_0$
  - $d(x_a, \mu_0) = \frac{d(\mu_a, \mu_0)}{1+\varepsilon}$
  - $d(x_a, y_a) = \frac{d(x_a, \mu_0)}{1+\varepsilon} = \frac{d(\mu_a, \mu_0)}{(1+\varepsilon)^2}$
  - Note that we can always choose such thresholds due to the continuity of KL-divergence.

- $\mathcal{H}_t = \{A_1, R_1, \cdots, A_{t-1}, R_{t-1}\}$ history of actions and rewards *before* time t.
  - For consistency, we use t for the subscript instead of $t-1$.

- We introduce two events.
  - $E^{\mu}_{a,t} = \{\hat{\mu}_{a,t-1} \le x_a\}$
  - $E^{\theta}_{a,t} = \{\theta_{a,t} \le y_a\}$
  - The idea is that these two events are more likely to happen as time progresses.
  - Note that given $\mathcal{H}_t$, $E^{\mu}_{a,t}$ is determined to be either true of false and $E^{\theta}_{a,t}$ is still random, but the probability that it happens can be exactly calculated.

- $p_{a,t} := \mathbb{P}(\theta_{0,t} > y_a | \mathcal{H}_t)$

1.3 Key Lemma

Lemma 2. *For all t and $a \neq 0$,*

$$\mathbb{P}(A_t = a, E^{\mu}_{a,t}, E^{\theta}_{a,t} | \mathcal{H}_t) \le \frac{1 - p_{a,t}}{p_{a,t}} \mathbb{P}(A_t = 0, E^{\mu}_{a,t}, E^{\theta}_{a,t} | \mathcal{H}_t)$$

Note: We expect $\frac{1-p_{a,t}}{p_{a,t}}$ to be very small.

Proof. Note that $E^{\mu}_{a,t}$ is fixed given $\mathcal{H}_t$ and there is nothing to prove when it does not happen. From now on, we will assume this event happened. Using Bayes rule, we get:

$$\begin{aligned}
\text{LHS} &= \mathbb{P}(A_t = a, E^{\theta}_{a,t} | \mathcal{H}_t) \\
&= \mathbb{P}(A_t = a | E^{\theta}_{a,t}, \mathcal{H}_t) \mathbb{P}(E^{\theta}_{a,t} | \mathcal{H}_t) \\
\text{RHS} &= \frac{1 - p_{a,t}}{p_{a,t}} \mathbb{P}(A_t = 0 | E^{\theta}_{a,t}, \mathcal{H}_t) \mathbb{P}(E^{\theta}_{a,t} | \mathcal{H}_t).
\end{aligned}$$

Thus it suffices to prove that $\mathbb{P}(A_t = a | E^{\theta}_{a,t}, \mathcal{H}_t) \le \frac{1-p_{a,t}}{p_{a,t}} \mathbb{P}(A_t = 0 | E^{\theta}_{a,t}, \mathcal{H}_t)$. Given $E^{\theta}_{a,t}$, $\{A_t = a\}$ implies that $\theta_{\tilde{a},t} \le y_a$ for any $\tilde{a} \in \mathcal{A}$. Therefore,

$$\begin{aligned}
\text{LHS} &\le \mathbb{P}(\forall \tilde{a} \in \mathcal{A}, \theta_{\tilde{a},t} \le y_a | E^{\theta}_{a,t}, \mathcal{H}_t) \\
&= \mathbb{P}(\theta_{0,t} \le y_a | E^{\theta}_{a,t}, \mathcal{H}_t) \mathbb{P}(\forall \tilde{a} \neq 0, \theta_{\tilde{a},t} \le y_a | E^{\theta}_{a,t}, \mathcal{H}_t) \\
&= (1 - p_{a,t}) \mathbb{P}(\forall \tilde{a} \neq 0, \theta_{\tilde{a},t} \le y_a | E^{\theta}_{a,t}, \mathcal{H}_t)
\end{aligned}$$

The last equality holds because the event $\{\theta_{a,t} \le y_a\}$ is independent of $E^{\theta}_{a,t}$. Similarly, given $E^{\theta}_{a,t}$, $\{A_t = 0\}$ happens whenever $\theta_{0,t} > y_a$ and $\theta_{\tilde{a},t} \le y_a$ for any $\tilde{a} \neq 0$. This gives us

$$\begin{aligned}
\text{RHS} &\ge \frac{1 - p_{a,t}}{p_{a,t}} \mathbb{P}(\theta_{0,t} > y_a, \theta_{\tilde{a},t} \le y_a \forall \tilde{a} \neq 0 | E^{\theta}_{a,t}, \mathcal{H}_t) \\
&= (1 - p_{a,t}) \mathbb{P}(\forall \tilde{a} \neq 0, \theta_{\tilde{a},t} \le y_a | E^{\theta}_{a,t}, \mathcal{H}_t)
\end{aligned}$$

The above two inequalities complete the proof of the key lemma. □

1.4 (Incomplete) Proof of Theorem

We will start by rewriting $N_T(a)$ as the sum of indicator variables and taking expectation.

$$\begin{aligned}
\mathbb{E}N_T(a) &= \sum_{t=1}^{T} \mathbb{P}(A_t = a) \\
&= \sum_{t=1}^{T} [\mathbb{P}(A_t = a, E^{\mu}_{a,t}, E^{\theta}_{a,t}) + \mathbb{P}(A_t = a, E^{\mu}_{a,t}, \overline{E^{\theta}_{a,t}}) + \mathbb{P}(A_t = a, \overline{E^{\mu}_{a,t}})] \\
&= S_1 + S_2 + S_3
\end{aligned}$$

In this (incomplete) proof, we will focus on S_1 only. We will control S_2 and S_2 in the next lecture. To impose an upper bound, we will use the tower property of expectation several times.

$$\begin{aligned}
\mathbb{P}(A_t = a, E^{\mu}_{a,t}, E^{\theta}_{a,t}) &= \mathbb{E}[\mathbb{P}(A_t = a, E^{\mu}_{a,t}, E^{\theta}_{a,t} | \mathcal{H}_t)] (\because \text{tower property}) \\
&\le \mathbb{E}[\frac{1 - p_{a,t}}{p_{a,t}} \mathbb{P}(A_t = 0, E^{\mu}_{a,t}, E^{\theta}_{a,t} | \mathcal{H}_t)] (\because \text{key lemma}) \\
&= \mathbb{E}[\mathbb{E}[\frac{1 - p_{a,t}}{p_{a,t}} \mathbb{I}(A_t = 0, E^{\mu}_{a,t}, E^{\theta}_{a,t} | \mathcal{H}_t)]] (\because p_{a,t} \text{ is a function of } \mathcal{H}_t) \\
&= \mathbb{E}[\frac{1 - p_{a,t}}{p_{a,t}} \mathbb{I}(A_t = 0, E^{\mu}_{a,t}, E^{\theta}_{a,t})] (\because \text{tower property}) \\
&\le \mathbb{E}[\frac{1 - p_{a,t}}{p_{a,t}} \mathbb{I}(A_t = 0)]
\end{aligned}$$

Let τ_k be the time at which action 0 is taken for the k^{th} time with $\tau_0 := 0$. The key idea is that $p_{a,t}$ is updated only when $\{A_t = 0\}$ happens.

$$\begin{aligned}
S_1 &\le \sum_{t=1}^{T} \mathbb{E}[\frac{1 - p_{a,t}}{p_{a,t}} \mathbb{I}(A_t = 0)] \\
&\le \mathbb{E}[\sum_{k=0}^{T-1} \frac{1 - p_{a,\tau_k+1}}{p_{a,\tau_k+1}} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}[A_t = 0]] \\
&= \mathbb{E}[\sum_{k=0}^{T-1} \frac{1 - p_{a,\tau_k+1}}{p_{a,\tau_k+1}}]
\end{aligned}$$

The second inequality might not be equality because k can be strictly less than $T-1$. The last equality holds because the second summation from the second line is exactly 1 by definition of τ_k.

The lecture ended by stating the following lemma without proof. The proof requires extensive calculations involving Beta and Binomial distributions. See the appendix of [AG13] for details.

Lemma 3.

$$\mathbb{E}[\frac{1-p_{a,\tau_k+1}}{p_{a,\tau_k+1}}] \leq \begin{cases} \frac{3}{\Delta'_a} & \textit{if } k < \frac{8}{\Delta'_a} \\ \Theta(e^{-k\Delta'^2_a/2} + \frac{1}{(k+1)\Delta'^2_a}e^{-kD_a} + \frac{1}{e^{k\Delta'^2_a/4}-1}) & \textit{otherwise} \end{cases}$$

where $\Delta'_a = \mu_0 - y_a$ *and* $D_a = d(y_a, \mu_0)$.

References

[AG12] Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1, 2012.

[AG13] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *AISTATS*, pages 99–107, 2013.

[Tho33] William 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.