IEOR 8100-001: Learning and Optimization for Sequential Decision Making 02/03/16

Lecture 5: Thomposon Sampling (part II): Regret bounds proofs

Instructor: Shipra Agrawal Scribed by: Ni Ma

In this lecture, we provide proof for regret bounds for Thompson Sampling with Gaussian priors, stated as Theorem 4 in Lecture 4.

1 Proof for two-armed case ($N = 2$)

Consider the special case of two arms. The mean rewards of arm 1 and arm 2 are μ_1 and μ_2 respectively. Without loss of generality, we assume that the first arm is the optimal arm $\mu_1 > \mu_2$ and $\Delta := \mu_1 - \mu_2$. We prove that in this case $E[n_{2,T}] \leq O(\frac{\log(T)}{\Delta^2})$, which would imply that regret is bounded by $E[R(T)] = E[n_{2,T}\Delta] = O(\frac{\log(T)}{\Delta})$.

1.1 Challenges

We describe the main technical difficulties in the proof for TS algorithm as compared to the UCB algorithm. In UCB algorithm, the suboptimal arm 2 will be played at time t, if its UCB value is higher, i.e. if $UCB_{2,t-1} > UCB_{1,t-1}$. If we have pulled arm 2 for some amount of times $\Omega(\frac{log(T)}{\Delta^2})$, then with a high probability this will not happen. This is because after $n_{2,t} \geq \Omega(\log(T)/\Delta^2)$, using concentration bounds we can derive that $UCB_{2,t}$ will be close to its true mean μ_2. So that, with high probability.

$$UCB_{2,t} \leq \mu_2 + \Delta \leq \mu_1 \leq UCB_{1,t}$$

The last inequality holds because UCB was defined so that it is always above the true mean for any arm.

In the TS algorithm, we generate a sample $\theta_{1,t}$ and $\theta_{2,t}$ and pull the arm i with a larger $\theta_{i,t}$. After some amount of pulls of arm 2, $\Omega(\frac{log(T)}{\Delta^2})$ as in UCB algorihm, $\theta_{2,t}$ will be concentrated around its true mean $\mu_2 \leq \mu_1$, which is not hard to prove. However, this is not sufficient because unlike UCB, $\theta_{1,t}$ is no longer guaranteed to be above μ_1. Rather, we need to wait until arm 1 has been played a moderate amount of times so that θ_1 is also concentrated around μ_1 with a large probability.

To summarize, in the proof for UCB algorithm, we wait until arm 2 has been played some amount of times and we are done. While in the proof for TS algorithm, we first wait for some amount of plays of arm 2, then wait for some amount of plays of arm 1, and afterwards we will always play the optimal arm 1 with a large probability.

1.2 Proof outline

To capture above intuition, in the analysis we divide the time horizon T into three phases.

- Phase 1. From the beginning of the time horizon until time t when arm 2 has had at least $\frac{64\log(T)}{\Delta^2}$ pulls, i.e. $n_{2,t} \geq \frac{64\log(T)}{\Delta^2}$.
- Phase 2. From the end of phase 1 until arm 1 has at least $\frac{64log(T)}{\Delta^2}$ pulls, i.e. $n_{1,t} \geq \frac{64\log(T)}{\Delta^2}$.
- Phase 3. From the end of phase 2 until the end of time horizon T.

To bound $E[n_{2,T}]$, we need to bound the number of pulls of arm 2 in each of these three phases. In phase 1, it is already bounded by $\frac{64\log(T)}{\Delta^2}$. It will be easy for phase 3, because we can show that arm 1 will be played with a probability of $1 - O(\frac{1}{T^2})$. Then the key step is to get the bound for phase 2.

1.3 Bounding regret in phase 2

For any time step t in phase 2, we have

$$n_{2,t-1} \geq \frac{64\log(T)}{\Delta^2} =: L. \tag{1}$$

First we observe that $\hat{\mu}_{2,t}$ is concentrated around μ_2, and $\theta_{2,t}$ is concentrated around $\hat{\mu}_{2,t}$ for all t.

Definition 1. *We say event $\mathcal{E}_t$ holds, if both of the following inequalities are true:*

$$|\hat{\mu}_{2,t-1} - \mu_2| \quad \leq \quad \sqrt{\frac{\log(T)}{n_{2,t-1}}}, \tag{2}$$

$$|\theta_{2,t} - \hat{\mu}_{2,t-1}| \quad < \quad \sqrt{\frac{4\log(T)}{n_{2,t-1}+1}}. \tag{3}$$

Lemma 2. *For any $t \geq 1$, $\Pr(\mathcal{E}_t) \geq 1 - \frac{3}{T^2}$.*

Proof. For any t, we have with probability $1 - \frac{2}{T^2}$:

$$|\hat{\mu}_{2,t-1} - \mu_2| \leq \sqrt{\frac{\log(T)}{n_{2,t-1}}}. \tag{4}$$

This follows easily from Chernoff-Hoeffding bounds. For concentration of $\theta_{2,t}$, we use the fact that for a Gaussian distributed random variable X with mean m and variance σ^2, for any z,

$$Pr(|X - m| > z\sigma) \geq 1 - \frac{1}{2}e^{-\frac{z^2}{2}}. \tag{5}$$

Now, since $\theta_{2,t} \sim \mathcal{N}(\hat{\mu}_{2,t-1}, \frac{1}{n_{2,t-1}+1})$, we have with probability $1 - \frac{1}{2T^2}$,

$$|\theta_{2,t} - \hat{\mu}_{2,t-1}| \quad \leq \quad \sqrt{\frac{4\log(T)}{n_{2,t-1}+1}} \tag{6}$$

Combining these two observations we get the lemma statement. □

Corollary 3. *Assume $\mathcal{E}_t$ holds at time t, and $n_{2,t} \geq L$ (i.e., t is in phase 2 or phase 3). Then, $\theta_{2,t} \leq \mu_1 - \frac{\Delta}{2}$.*

Proof. We substitute $n_{2,T} = L = \frac{64\log(T)}{\Delta^2}$ in Equation (2) and Equation (3), to get

$$\theta_{2,t} \leq \hat{\mu}_{2,t-1} + \frac{\Delta}{4} \leq \mu_2 + \frac{\Delta}{4} + \frac{\Delta}{4} = \mu_1 - \frac{\Delta}{2}. \tag{7}$$

Here, the second inequality is loose. □

Above corollary implies that if we also have $\theta_{1,t} \geq \mu_1 - \frac{\Delta}{2}$ with a large probability, then we are done – $\theta_{1,t}$ would be greater than $\theta_{2,t}$ with high probability. We will have this with high probability in phase 3, i.e., once arm 1 has been pulled L times as well. Until then, we show that this happens (in expectation) every constant number of time steps. And, that will be enough to bound the number of pulls in phase 2.

Definition 4. *For any time t, define p_t as the probability that $\theta_{1,t}$ exceeds μ_1:*

$$p_t := \Pr(\theta_{1,t} \geq \mu_1).$$

Note that p_t is determined by the distribution of $\theta_{1,t}$, which is determined by history $\mathcal{H}_{t-1}$ before time t.

Note two further properties of quantity p_t as defined above, which will be useful later in the proof.

- First, the event $\mathcal{E}_t$ is about sample outcomes and reward observations of plays of arm 2, which has no implications on the sample outcomes of plays of arm 1. Then the event $\{\theta_{1,t} \geq \mu_1\}$ is independent of $\mathcal{E}_t$ and thus
$$p_t := \Pr(\theta_{1,t} \geq \mu_1) = \Pr(\theta_{1,t} \geq \mu_1 | \mathcal{E}_t). \tag{8}$$
- Second, we only update the distribution of $\theta_{1,t}$ when arm 1 is played, so p_t changes value only when arm 1 is played. Thus, $p_s = p_t$ if no pulls of arm 1 happened between from time s to time $t-1$.

Now, we are ready to prove the central lemma in this proof, which we call "Phase 2 lemma". This lemma upper bounds the probability of playing arm 2 in terms of probability of playing arm 1.

Lemma 5. *(Phase 2 Lemma) For any time t in phase 2,*
$$Pr(I_t = 2|\mathcal{E}_t) \leq \left(\frac{1}{p_t} - 1\right) Pr(I_t = 1|\mathcal{E}_t). \tag{9}$$

Proof. Using Corollary 3, we have that for any time t in phase 2 (i.e., for t such that $n_{i,t-1} \geq L$),
$$\begin{aligned} \Pr(I_t = 1|\mathcal{E}_t) &= \Pr(\theta_{1,t} > \theta_{2,t}|\mathcal{E}_t) \\ &\geq Pr(\theta_{1,t} \geq \mu_1|\mathcal{E}_t) = p_t. \end{aligned}$$
Now, since $\Pr(I_t = 1|\mathcal{E}_t) + \Pr(I_t = 2|\mathcal{E}_t) = 1$, we get
$$\Pr(I_t = 2|\mathcal{E}_t) \quad \leq \quad 1 - p_t.$$
The ratio of these two inequalities gives the desired result. □

We state following bound on $E[\frac{1}{p_t} - 1]$ without formal proof. Refer to [1] for a complete proof.

Lemma 6. *For any t,*
$$E[\frac{1}{p_t} - 1] \leq (e^{11} + 5). \tag{10}$$

Proof intuition. Note that the lemma statement is required to hold irrespective of the number of plays $n_{1,t-1}$ of arm 1 at time t. If the number of plays of arm 1 is small, then the empirical mean $\hat{\mu}_{1,t-1}$ could be far away from actual mean μ_1 and possibly much smaller than μ_1, making it difficult for $\theta_{1,t}$ to exceed μ_1 ($\theta_{1,t}$ is Gaussian with mean $\hat{\mu}_{1,t-1}$). However in that case, the variance of $\theta_{1,t}$ is also high ($\theta_{1,t}$ is Gaussian with variance $\frac{1}{n_{1,t-1}+1}$), and therefore by anti-concentration properties of Gaussian, there is significant probability of $\theta_{1,t}$ to exceed its mean enough to exceed μ_1. The proof of this lemma is achieved by a careful balancing between concentration of empirical mean and anti-concentration of Gaussian. □

Intuitively, the combination of above two lemmas says that we will see a play of arm 1 after every few (constant) plays of arm 2. Therefore, if we wait until L plays of arm 1, we will not see more than a constant times L plays of arm 2. More formally, we can use the results above to bound the number of plays of arm 2 in phase 2 as follows. Let τ_1, τ_2 denote the beginning and end time steps of phase 2. Recall that phase 2 was defined to end as soon as

$n_{1,t}$ exceeds $L = \frac{64\log(T)}{\Delta^2}$. Let $\bar{\mathcal{E}}_t$ denote the complement of the event $\mathcal{E}_t$.

$$
\begin{aligned}
E[\sum_{t=\tau_1}^{\tau_2} \mathbb{1}(I_t = 2)] &\leq E[\sum_{t=\tau_1}^{\tau_2} [Pr(I_t = 2|\mathcal{E}_t)\Pr(\mathcal{E}_t) + \Pr(\bar{\mathcal{E}}_t, I_t = 2)]] && (11)\\
(*) &\leq E[\sum_{t=\tau_1}^{\tau_2} \Pr(I_t = 2|\mathcal{E}_t)\Pr(\mathcal{E}_t)] + \sum_{t=1}^{T}\Pr(\bar{\mathcal{E}}_t) && (12)\\
&\leq E[\sum_{t=\tau_1}^{\tau_2} \frac{(1-p_t)}{p_t}\Pr(I_t = 1|\mathcal{E}_t)\Pr(\mathcal{E}_t)] + \frac{3}{T} && (13)\\
&= E[\sum_{t=\tau_1}^{\tau_2} \frac{(1-p_t)}{p_t}\Pr(I_t = 1, \mathcal{E}_t)] + \frac{3}{T} && (14)\\
&\leq E[\sum_{t=\tau_1}^{\tau_2} \frac{(1-p_t)}{p_t}\Pr(I_t = 1)] + \frac{3}{T} && (15)\\
&\leq E[\sum_{t=\tau_1}^{\tau_2} \frac{(1-p_t)}{p_t} I(I_t = 1)] + \frac{3}{T} && (16)\\
&\quad \text{(let } \gamma_j \text{ denotes the time step of the } j\text{th play of arm 1.)} && (17)\\
(**) &\leq E[\sum_{j=j_0}^{L} \frac{(1-p_{\gamma_j})}{p_{\gamma_j}}] + \frac{3}{T} && (18)\\
&\leq constant \times L + \frac{3}{T}. && (19)
\end{aligned}
$$

(*) follows from the phase 2 lemma, and Lemma 2. In (**), we used that arm 1 cannot be played more than L time in phase 2, because phase 2 is ended after that. Here, j_0 denotes the number of plays of arm 1 at the beginning of phase 2. The last inequality followed from the bound in Lemma 6.

1.4 Bounding regret in Phase 3

In phase 3, both arms are concentrated well enough. Similar to proof of Lemma 2, we obtain that with probability $1 - \frac{6}{T^2}$,

$$
\begin{aligned}
|\hat{\mu}_{1,t-1} - \mu_2| &\leq \sqrt{\frac{\log(T)}{n_{1,t-1}}} \leq \frac{\Delta}{4} && (20)\\
|\hat{\mu}_{2,t-1} - \mu_2| &\leq \sqrt{\frac{\log(T)}{n_{2,t-1}}} \leq \frac{\Delta}{4} && (21)\\
\theta_{2,t} \leq \hat{\mu}_{2,t-1} + \sqrt{\frac{4\log(T)}{n_{2,t-1}}} &\leq \hat{\mu}_{2,t-1} + \frac{\Delta}{4} \leq \mu_2 + \frac{\Delta}{2} && (22)\\
\theta_{1,t} \geq \hat{\mu}_{1,t-1} - \sqrt{\frac{4\log(T)}{n_{1,t-1}}} &\geq \hat{\mu}_{1,t-1} - \frac{\Delta}{4} \geq \mu_1 - \frac{\Delta}{2}. && (23)
\end{aligned}
$$

Thus $\theta_{1,t} \geq \theta_{2,t}$ with probability $1 - \frac{6}{T^2}$. Therefore, in phase 3, expected number of plays of arm 2 is bounded by $\frac{6}{T}$.

Combining the observations for the three phases, we obtain that

$$E[n_{2,T}] \le \underbrace{\frac{64\log(T)}{\Delta^2}}_{\text{phase 1}} + \underbrace{(e^1 1+5)\frac{64\log(T)}{\Delta^2} + \frac{3}{T}}_{\text{phase 2}} + \underbrace{\frac{6}{T}}_{\text{Phase 3}} = O(\frac{\log(T)}{\Delta^2})$$

2 Multiple-Armed Case

The intuition behind analyzing TS algorithm (and proving Theorem 4) for the general N-armed case is same as 2-armed case. W.l.o.g., again assume that arm 1 is optimal arm, and for every other arm $i \neq 1$, $\mu_i < \mu_1$, $\Delta_i := \mu_1 - \mu_i$. Now, for every suboptimal arm i, we bound the number of plays $E[n_{i,T}]$ separately by dividing the time horizon into three phases. When boudning plays of arm i,

- Phase 1 is defined from beginning until the time arm i has had $L_i := \frac{64\log(T)}{\Delta_i^2}$ plays,
- Phase 2 is defined from end of phase 1 until arm 1 has had L_i plays.

Now define $\mathcal{E}_{i,t}$ is the event that $\hat{\mu}_{i,t}$ and $\theta_{i,t}$ follow concentration bounds. This event can be proven to hold with high probability (similar to Lemma 2), and given this event, in phase 2, $\theta_{i,t} \le \mu_1$ (similar to Corollary 3). Therefore, if $\theta_{1,t}$ exceeds μ_1, then we would have $\theta_{1,t} > \theta_{i,t}$. Now, the main difficulty compared to the two-armed case is that $\theta_{1,t} > \theta_{i,t}$ is not sufficient to ensure that arm 1 will be played. (For some other arm $j \neq i$, $\theta_{j,t}$ may exceed $\theta_{1,t}$). The key to handle this is the following "phase 2 lemma", which shows that this observation can still be used to upper bound the probability of playing arm i in terms of probability of playing arm 1.

Lemma 7. *(Multi-arm Phase 2 Lemma) For any t, such that $n_{i,t-1} \ge L_i := \frac{64\log(T)}{\Delta^2}$,*

$$Pr(I_t = i|\mathcal{E}_{i,t}, H_{t-1}) \le \frac{(1-p_t)}{p_t} Pr(I_t = 1|\mathcal{E}_{i,t}, H_{t-1}). \tag{24}$$

Proof. Given $n_{i,t-1} \ge L_i$, the event $\mathcal{E}_{i,t}$ implies that $\theta_{i,t} \le \mu_1$. Let $p_t := Pr(\theta_{1,t} \ge \mu_1|H_{t-1}) = Pr(\theta_{1,t} \ge \mu_1|\mathcal{E}_{i,t}, H_{t-1})$, where the latter equality follows from observing that $\theta_{1,t}$ is independent of $\theta_{i,t}$.

$$\begin{aligned}
\Pr(I_t = i|\mathcal{E}_{i,t}, H_{t-1}) &\le \Pr(\theta_{j,t} \le \mu_1, \forall j \neq i) \\
&= \Pr(\theta_{j,t} \le \mu_1, \forall j \neq i, j \neq 1|H_{t-1}, \mathcal{E}_t) \times \Pr(\theta_{1,t} \le \mu_1|\mathcal{E}_{i,t}, H_{t-1}) \\
&= Pr(\theta_{j,t} \le \mu_1, \forall j \neq i, j \neq 1|H_{t-1}, \mathcal{E}_t) \times (1-p_t) \\
Pr(I_t = i|\mathcal{E}_t, H_{t-1}) &\ge Pr(\theta_{j,t} \le \mu_1, \forall j \neq i, j \neq 1|H_{t-1}, \mathcal{E}_t) \times Pr(\theta_{1,t} \ge \mu_1|\mathcal{E}_{i,t}, H_{t-1}) \\
&= Pr(\theta_{j,t} \le \mu_1, \forall j \neq i, j \neq 1|H_{t-1}, \mathcal{E}_t) \times p_t.
\end{aligned}$$

The ratio of these two inequalities gives the result. □

Remaining steps to bound number of pulls of arm i in phase 2 are similar to the Equations (11)-(19). Refer to [1] for complete proof.

References

[1] S. Agrawal, N. Goyal, "Further optimal regret bounds for Thompson Sampling", In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), 2013.