

Lecture 18 — October 7

Lecturer: Aditya Gopalan

Scribe: Praveen M P

18.1 Recap

In the last class we looked at the Upper Confidence Bound (UCB) algorithm for minimizing the regret bound in the Stochastic Multi-Armed Bandit setting. We saw that UCB artificially introduces a bias to the empirical mean of rewards, which is used to choose best performing arm so far. We also proved the regret bound performance of UCB algorithm for $[0,1]$ rewards, which turned out to be:

$$\text{Regret}_T^{UCB} \leq \delta \log(T) \left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right) + \frac{\pi^2}{3} \sum_{i=1}^N \Delta_i \quad \text{where, } \Delta_i = \mu^* - \mu_i, \mu^* = \max_i \mu_i$$

We also looked at a different setting of ‘pure exploration in stochastic bandits’, to identify the best arm as quickly as possible. After defining the notion of (ϵ, δ) -PAC algorithm with sample complexity T , we studied two (ϵ, δ) -PAC algorithms, the naive algorithm, and median elimination algorithm, for which the sample complexities was shown to be $\mathcal{O}\left(\frac{2N}{\epsilon^2} \log\left(\frac{2N}{\delta}\right)\right)$ and $\mathcal{O}\left(\frac{N}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right)$.

In this class we will look at an entirely different algorithmic framework for approaching the S-MAB problem, called the Thompson sampling. We will also do the regret bound analysis of the algorithm.

18.2 Thompson sampling for the S-MAB problem

Thompson sampling[1] is a classical algorithm, that has been in use since decades for addressing the exploration-exploitation dilemma in the stochastic multi-armed bandit problem, *i.e* we want to explore for the best arm, but at the mean time we also want to minimise the regret incurred in the process. Even though the algorithm worked well, and even better than the UCB method, it remained as a heuristic, and no analysis was done for the regret bounds of the algorithm, till 2011. We will look at the Thompson sampling in detail, and also look at its regret bound that was proved in [2].

Consider an N -armed S-MAB setting, with bandits having bernoulli rewards $X_{i,s} \sim \text{Ber}(\theta_i)$, $\theta_i \in [0, 1]$, independent in $i \in [N]$ and iid in s for each i . The basic idea of thompson sampling is to give a bayesian interpretation to the S-MAB problem. The bayesian view of S-MAB problem illustrated in Algorithm 1, and the Thompson sampling algorithm is given in Algorithm 2.

Algorithm 1 Bayesian inspiration of SMAB

Initialization: Start with a "belief" or "prior" distribution for each $\theta_i, i \in [N]$, say Unif [0,1]

At each t=1,2...T

-Sample $\theta_i(t) \sim \text{belief} \forall i \in [N]$

-Play the arm $I_t = \underset{i \in [N]}{\operatorname{argmax}} \theta_i(t)$

-For the played arm, update the Posterior belief(θ) \propto Prior(θ) $\cdot \mathbb{P}(\text{reward}/\theta)$

end

Algorithm 2 Thompson Sampling

At time t, let arm i sees $S_i(t)$ successes (reward=1) and $F_i(t)$ failures (reward=0)

Initialization: $S_i=0, F_i=0 \forall i \in [N]$

At each t=1,2...T do

-Sample $\theta_i(t) \sim \text{Beta}(1 + S_i, 1 + F_i)$ independently $\forall i \in [N]$

-Play the arm $I_t = \underset{i \in [N]}{\operatorname{argmax}} \theta_i(t)$

-Get the reward R_t for playing the arm I_t ; $R_t \in \{0, 1\}$

-Update

$S_{I_t} \leftarrow S_{I_t} + \mathbb{I}_{R_t=1}$

$F_{I_t} \leftarrow F_{I_t} + \mathbb{I}_{R_t=0}$

end

Notes on Thompson sampling:

- History : Thompson sampling [aka Posterior matching] [1]
 - It remained as heuristic for almost 80 years
 - First theoretical analysis in 2010 : showed regret bound of $\mathcal{O}(T)$ regret
 - Rigorous regret analysis done by Agrawal and Goyal in 2011 [2]
- Thompson sampling (TS) is very different from UCB style algorithms
 - TS is a randomized algorithm, given history (UCB, in contrast is deterministic)
 - The exploration is provided by sampling from the posterior. In fact, without random sampling, it is hopeless, because :
 - $\tilde{\theta}_i(t) = \mathbb{E} [\text{Beta}(1 + S_i, 1 + F_i)] \approx \frac{S_i}{S_i + F_i} = \frac{S_i}{T_i(t)}$ is like FTL, which gives bad regrets.
- Intuition
 - As an arm is played more and more $\text{Beta}(1+S_i, 1+F_i) \xrightarrow{d} \delta(\theta_i)$ where θ_i is the true mean for reward of arm i.

Before going to the analysis of Thompson sampling, lets review about the beta distribution and its properties.

18.2.1 Beta(a,b) distribution

$$f_{\text{Beta}(a,b)}(z) = \begin{cases} \frac{z^{a-1}(1-z)^{b-1}}{\int_0^1 y^{a-1}(1-y)^{b-1} dy} & \text{if } z \in [0, 1] \\ 0 & \text{if } z \notin [0, 1] \end{cases} \quad (18.1)$$

Notes:

- $\text{Beta}(1,1) \equiv U[0,1]$ • $\mathbb{E}[\text{Beta}(1,1)] = \frac{a}{a+b}$
- Beta distribution family is a conjugate prior family for the Bernoulli distribution.

i.e. Suppose we have $\theta \sim \text{Beta}(a,b)$ (Prior) and $X \sim \text{Ber}(\theta)$ (Likelihood Distribution), then $\theta/X \sim \text{Beta}(g(a,b,x))$ (Posterior)

More precisely,

$$\mathbb{P}[\theta \in A/X] = \int_A f_{\text{Beta}(a+x,b+1-x)}(z) dz$$

Likelihood Distribution(LD)	Parameter of LD	Conjugate prior family
Bernoulli	p(Probability of success)	Beta
Poisson	λ (Rate)	Gamma
Categorical(N)	$p \in \Delta_N$	Dirichlet $\in \Delta_N$
Gaussian(fixed variance)	μ (mean)	Gaussian
Gaussian(fixed mean)	σ^2 (Variance)	Inverse Gamma
Exponential distribution(λ)	λ	Gamma

Table 18.1. Likelihood distributions and their conjugate prior families

18.3 Illustrative example of TS performance : 2 Armed Bandit

Consider a 2-armed bernoulli bandit model, with mean rewards, $1 > \mu_1 = \mu^* > \mu_2 > 0$ and let $\Delta = \mu_1 - \mu_2$.

Theorem 18.1 (Agrawal, Goyal '11 ,“ Analysis of TS for the MAB problem ”).

$$\text{Expected Regret}_T^{TS} = \mathcal{O}\left(\frac{\log T}{\Delta} + \frac{1}{\Delta^3}\right) \quad (18.2)$$

Proof: To bound the expected regret, we will bound $\mathbb{E}[T_2(t)]$, the expected number of times sub-optimal arm is played, and multiplying this by Δ to get expected regret.

$$\text{Let } L = \frac{24 \log(T)}{\Delta^2}$$

Lets define random variables :

$J_0 :=$ Number of plays of arm 1, until arm 2 is played L times.

$V_j :=$ Time step at which arm 1 is played for the j^{th} time.

$Y_j := V_{j+1} - V_j - 1 =$ Number of time steps between the j^{th} and $(j+1)^{th}$ plays of arm 1

$S_{1,j} :=$ Number of 1's in the first j plays of arm 1.

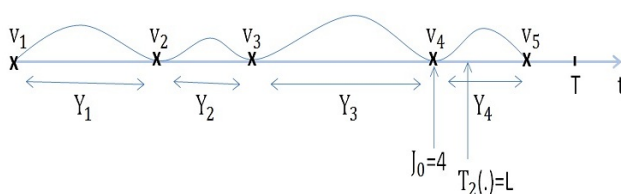


Figure 18.1.

Now, from the figure 18.1, its clear that,

$$\mathbb{E}[T_2(T)] \leq L + \mathbb{E} \left[\sum_{j=J_0}^{T-1} Y_j \right] \quad (18.3)$$

Lets define a random variable $W(j,s,y)$ as follows :

- Take repeated iid samples of a $\text{Beta}(1+s, 1+j-s)$ Random variable, and $W(\cdot)$ is the number of samples before the random variable exceeds y .
- $W(j,s,y)$ is a geometric random variable with success probability $= 1 - F_{\text{Beta}(1+s, 1+j-s)}(y)$.

Recall that

- $Y_j =$ Number of steps before $\{\theta_1(t) > \theta_2(t)\}$ happens for the first time after arm 1 is played for the j^{th} time.
- $W(j, S_{1,j}, y) =$ Number of steps before $\{\theta_1(t) > y\}$ happens for the first time after arm 1 is played for the j^{th} time.

Lets define an event $E = \{\forall t \in V_j + 1, \dots, V_{j+1} - 1, \theta_2(t) \leq \mu_2 + \frac{\Delta}{2}\}$

Now,

$$\mathbb{E}[Y_j \mathbb{I}_E] \leq \mathbb{E} \left[W(j, S_{1,j}, \mu_2 + \frac{\Delta}{2}) \wedge T \right] \quad (18.4)$$

where $x \wedge y = \min\{x, y\}$

Inequality 18.4 follows because, under the event E , $\mu_2 + \frac{\Delta}{2} \geq \theta_2(t)$, and thus from the definition of W , it is evident that Y_j , which is the number of plays until the fixed $\theta_1(t)$ is more than $\theta_2(t)$ which is upper bounded by $\mu_2 + \frac{\Delta}{2}$, is a 'probabilistically smaller' random variable than $W(j, S_{1,j}, \mu_2 + \frac{\Delta}{2})$

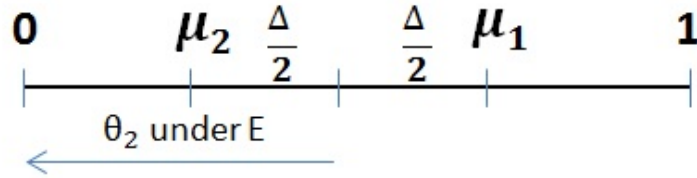


Figure 18.2. An illustration of location of θ_1 and θ_2

Thus,

$$\begin{aligned} \mathbb{E}[Y_j] &\leq \mathbb{E}\left[W(j, S_{1,j}, \mu_2 + \frac{\Delta}{2}) \wedge T\right] + \mathbb{E}[T\mathbb{I}_{E^c}] \\ &\leq \mathbb{E}\left[\underbrace{W(j, S_{1,j}, \mu_2 + \frac{\Delta}{2})}_{W_j} \wedge T\right] + \mathbb{E}\left[T \sum_{t=V_j+1}^{V_{j+1}-1} \mathbb{I}_{\{\theta_2(t) > \theta_1(t)\}}\right] \end{aligned} \quad (18.5)$$

where, the first inequality follows from the fact that $\mathbb{P}(A) \leq \mathbb{P}(A, B) + \mathbb{P}(B^c)$
Summing 18.5 over $j=J_0$ to $T-1$,

$$\begin{aligned} \mathbb{E}\left[\sum_{j=J_0}^T Y_j\right] &= \mathbb{E}\left[\sum_{j=0}^{T-1} Y_j \mathbb{I}_{j \geq J_0}\right] \\ &\leq \sum_{j=J_0}^{T-1} \mathbb{E}[W_j \wedge T] + T \sum_{j=J_0}^T \mathbb{E}\left[\sum_{t=V_j+1}^{V_{j+1}-1} \mathbb{I}_{\{\theta_2(t) > y, j \geq J_0\}}\right] \\ &\leq \sum_{j=J_0}^{T-1} \mathbb{E}[W_j \wedge T] + T \sum_{j=J_0}^T \mathbb{P}[\theta_2(t) > y, T_2(t) > L] \end{aligned} \quad (18.6)$$

Define the event $\mathbb{E}_2(t) := \{\theta_2(t) \leq \mu_2 + \frac{\Delta}{2} \text{ OR } T_2(t) < L\}$. We expect this to be a high probability event.

Lemma 1:

$$\mathbb{P}\{E_2(t)\} \geq 1 - \frac{2}{T^2} \quad \forall t \in \{1, 2, \dots, T\} \quad (18.7)$$

Lemma 2:

$$\begin{aligned} \mathbb{E}[W(j, S_{1,j}, y) \wedge T] &= \mathbb{E}[\mathbb{E}[W(j, S_{1,j}, y) \wedge T / S_{1,j}]] \\ &\leq \begin{cases} 1 + \frac{2}{1-y} + \frac{\mu_1}{\Delta^j} e^{-D_j} & \text{if } j \leq \frac{y \log R}{D} \\ 1 + \frac{R^y e^{-D_j}}{1-y} + \frac{\mu_1}{\Delta^j} e^{-D_j} & \text{if } \frac{y \log R}{D} < j < \frac{4 \log T}{\Delta^2} \\ \frac{16}{T} & \text{if } j \geq \frac{4 \log T}{\Delta^2} \end{cases} \end{aligned} \quad (18.8)$$

Where $D = \text{KL Divergence}(y, \mu_1)$ and $R = \frac{\mu_1}{1-\mu_1} \left(\frac{y}{1-y} \right)$, $y = \mu_2 + \frac{\Delta}{2}$ and $\Delta' = \frac{\Delta}{2}$. The outer expectation is over $S_{1,j} \sim \text{Bin}(j, \mu_1)$ and the inner expectation is over randomness in \mathbf{W} , which is a geometric random variable.

Using 18.7 and 18.8 in 18.6 and substituting in 18.3 ,

$$\begin{aligned} \mathbb{E}[T_2(t)] &\leq L + \frac{4 \log T}{\Delta'^2} + \underbrace{\sum_{j=0}^{\frac{4 \log T}{\Delta'^2}} \frac{\mu_1 e^{-D_j}}{\Delta'} + \frac{y \log R}{D} \frac{2}{1-y} + \sum_{j=\frac{y \log R}{D}}^{\frac{4 \log T}{\Delta'^2}} \frac{R^y e^{-D_j}}{1-y}}_{O(1)} + \frac{16}{T} T + \frac{2}{T^2} T \cdot T \quad (1) \\ &\leq \frac{40 \log T}{\Delta^2} + \frac{48}{\Delta^4} + 18 \quad (18.9) \end{aligned}$$

Now, multiplying 18.9 by Δ , we get the expected regret as in 18.2

□

18.4 References

- [1] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, *Biometrika*, 1933
- [2] S. Agrawal, N. Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem, 2011.