E1 245: Online Prediction & Learning **Fall 2014**

Lecture 19 — Oct 9

Lecturer: Aditya Gopalan *Scribe: Ganesh Ghalme*

19.1 Thompson Sampling

19.1.1 Recap

Last time we studied the overview of Thompson sampling and how this analysis technique can be used effectively to bound regret in MAB problem.
For N armed bandit problem and Bernaulli reward assumption we have Thompson sampling algorithm,

Algorithm 1: Thompson Sampling

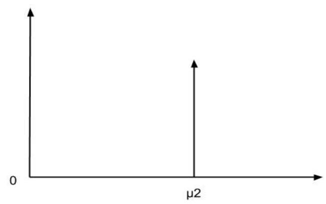
Input: No of arms = N, rewards dist=$Bern(\theta_i)$
Initialize: S_i =F_i =0 (S_i = No of successes i e 1's) $\forall i \in [N]$
At time t= 1,2,3,...
Sample $\theta_i(t) \sim Beta(1+S_i, 1+F_i)$
Play arm ,$I_t = \arg\max_i \theta_i(t)$, get reward R_t ;
Update
$S_{I_t} = S_{I_t} + \mathbb{1}(\mathbb{R}_T = 1)$
$F_{I_t} = F_{I_t} + \mathbb{1}(\mathbb{R}_T = 0)$

19.1.2 Two Arms case

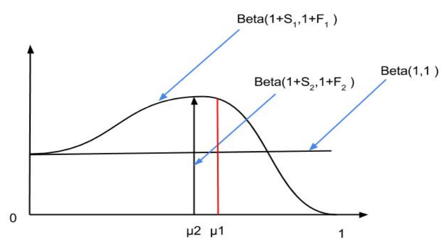
High level analysis Idea: Assume $\mu_1 \geq \mu_2$

Suppose arm 2 (sub-optimal arm) behaves ideally i.e. at any time t, $\theta_2(t) \simeq \mu_2$, which is equivalent to saying you have perfect information about arm 2.

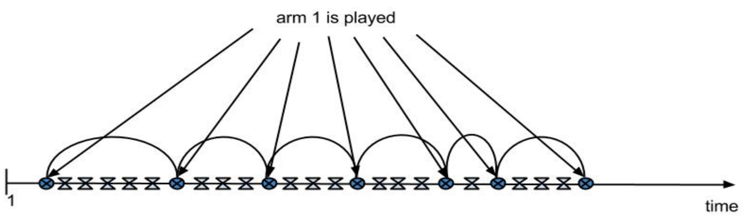
The posterior distribution of arm 2 looks like.



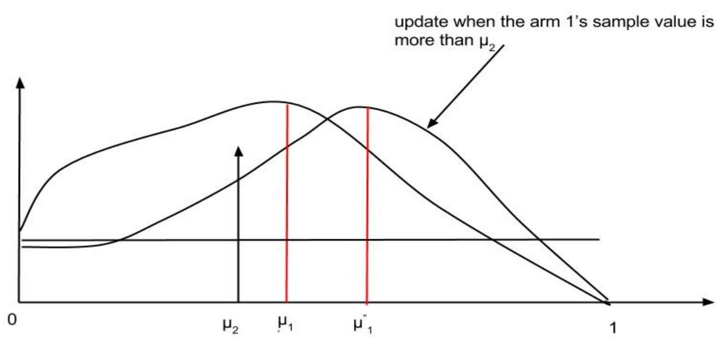
Posterior distribution of arm 1, initially an uniform distribution, would gradually look like,



The regret is incurred only when algorithm decides to play sub-optimal arm (i. e. arm 2)
Looking at the picture in time.



Whenever the sampled value $\theta_1(t)$ is greater than μ_2 posterior of arm 1 is updated.
Which is equivalent to saying we update confidence level of arm 1 only when we choose it.



Next we address the question that without assumption 1, how much do arm 1's posterior sample deviate from typicality?

19.1.3 Proof

Notations

1. J_0= Number of plays of arm 1 when arm 2 is played for $L = 24 * logT/\Delta^2$ number of times. [L is the time point at which assumption 1 kicks in]
2. j = Number of plays of arm 1 with S successes and (j-S) failures.
3. V_j= Time step at which j'th play of arm 1 happens.
4. $Y_j = V_{j+1} - V_j - 1$, Measure of time steps between j'th and (j+1)th play of arm 1.
5. $W(j,S,y)$= Number of trials of a $Beta(1+S, 1+j-S)$ distribution.where S is the number of successes.

After L round
Expected number of plays of second arm in time T is bounded by

$$\mathbb{E}(T_2(T)) \le L + \mathbb{E} \sum_{j=J_0}^{T-1} Y_j$$

To understand the expectation of Y_j we do following until it succeeds

1. Check if $Beta(1+S, 1+j-S)$ distributed random variable exceeds a threshold y.
2. For each experiment we generate a beta-distributed r.v. independently of previous ones.
3. $W(j,S,y)$ denotes the number of trials before the experiment succeeds. it takes non-negative integer values and is geometric random variable.

Recall that Y_j is defined as the number of steps before $\theta_1(t) > \theta_2(t)$ happens for the first time after the j^{th} play of the first arm.
Now consider the steps before $\theta_1(t) > \mu_2 + \Delta/2$ happens for the first time after the j^{th} play of the first arm.
Let us consider the event where value of $\theta_2(t)$ lies below $\mu_2 + \Delta/2$ i.e.,

$$E = \{\forall t \in V_j + 1, ..., V_{j+1} - 1, \theta_2(t) \le \mu_2 + \Delta/2\}$$

Lets try to find $\mathbb{E}[Y_j \mathbb{1}_E]$ i.e. Expected value of Y_j under E,which can be bounded as follows

$$\mathbb{E}[Y_j \mathbb{1}_E] \le \mathbb{E}[W(j, S_{1j}, \mu_2 + \Delta/2) \cap T]$$

When $\theta_2(t) > \mu_2 + \Delta/2$, then we use the fact that Y_j is always bounded by T. Using the fact that $\mathbb{P}(A) \le \mathbb{P}(A,B) + \mathbb{P}(B^C)$ we have,

$$\mathbb{E}(y_j \mathbb{1}_E) \le \mathbb{E}[W(j, S_{1j}, \mu_2 + \Delta/2) \cap T] + \mathbb{E}(T \mathbb{1}_{E^C}) \text{ where } E^C \le \sum_{t=V_j+1}^{V_{j+1}-1} \mathbb{1}(\theta_2(t) > \mu_2 + \Delta/2)$$

$$\mathbb{E}(\sum_{j=J_0}^{T} y_j) \le \sum_{j=0}^{T} \mathbb{E}(W(j, S_{1,j}, y) \cap T) + T \sum_{j=0}^{T} \mathbb{E}(\sum_{t=v_j+1}^{v_{j+1}-1} \mathbb{1}(\{\theta_2(t) \ge y, j \ge J_0\})) \tag{19.1}$$

Where, $y = \mu_2 + \Delta/2$

Define:

$$E_2(t) = \{\theta_2(t) \leq y \text{ OR } T_2(t) < L\}$$

we want to bound $P(E_2(t^C)) \leq \ldots$
Lemma 1 : Key lemma.

$$\forall t \leq T \ , P(E_2(t)) \geq 1 - \frac{2}{T^2}$$

Proof of lemma 1:
Two sources of randomness one from samples from beta, second by sequence of seen variables.
$(S_2, F_2) \to Beta(S_2, F_2) \to \theta_2$
What we want to analyze is,

$$\mathbb{P}(E_2(t)^C) = P(\theta_2(t) > \mu_2 + \Delta/2, T_2 \geq L)$$

Introduce an auxilliary event,

$$A(t) = \{\frac{S_2(t)}{T_2(t)} \leq \mu_2 + \frac{\Delta}{4}\}$$

Idea-

$$\theta_2 - \mu_2 = \{(\theta_2(t) - \frac{S_2(t)}{T_2(t)}) + (\frac{S_2(t)}{T_2(t)} - \mu_2)\}$$

Where first term is the beta distribution deviation and second term is empirical mean deviation.

now, We make use of the fact $\mathbb{P}(A) \leq \mathbb{P}(A, B) + \mathbb{P}(B^C)$ to get,

$$\mathbb{P}(E_2(t)^C) \leq \mathbb{P}(A(t)^C, T_2(t) \geq L) + \mathbb{P}(A(t), T_2(t) \geq L, \theta_2(t) \geq \mu_2 + \Delta/2) \tag{19.2}$$

Consider the first term,

$$\mathbb{P}(A(t)^C, T_2(t) \geq L) = \mathbb{P}(\frac{S_2(t)}{T_2(t)} > \mu_2 + \Delta/4, T_2(t) \geq L)$$

Define another random variable $X_{2,M}$ as the average number of successes over the first M plays of the second arm. More precisely, let random variable $X_{2,m}$ denote the output of the m^{th} play of the second arm. Then,

$$X_{2,M} = \frac{1}{l} \sum_{m=1}^{l} X_{2,m}$$

and $\frac{S_2(t)}{T_2(t)}$ is the unbiased estimate of $X_{2,M}$
Using above results we can write,

$$\sum_{l=L}^{T} \mathbb{P}(\frac{S_2(t)}{T_2(t)} > \mu_2 + \Delta/4, T_2(t) = l) \leq \sum_{l=L}^{T} \mathbb{P}(1/l \sum_{m=1}^{l} X_{2,m} > \mu_2 + \Delta/4)$$

Using Azuma hoeffdings inequality RHS can be upper bounded by,

$$RHS \leq \sum_{l=L}^{T} exp(-2l\Delta^2/16)$$

Which can further be upper bounded by taking the lowest value of l i.e. $l = L$

$$\leq T * exp(-2L\Delta^2/16)$$

Putting optimal value of L , i.e. $L = 24 * logT/\Delta^2$

$$\implies RHS \leq T * exp(\frac{-2\Delta^2}{16}.\frac{24 * logT}{\Delta^2}) = 1/T^2$$

Consider the second term

$$\mathbb{P}(A(t), T_2(t) \geq L, \theta_2(t) > \mu_2 + \frac{\Delta}{2}) = \sum_{l=L}^{T} \mathbb{P}(\frac{S_2(t)}{T_2(t)} \leq \mu_2 + \frac{\Delta}{4}, \theta_2(t) > \mu_2 + \frac{\Delta}{2}, T_2(t) = l)$$

$$\leq \sum_{l=1}^{T} \mathbb{P}(\theta_2(t) > \frac{S_2(t)}{T_2(t)} - \frac{\Delta}{4} + \frac{\Delta}{2}, T_2(t) = l)$$
$$= \sum_{l=1}^{T} \mathbb{P}(\theta_2(t) > \frac{S_2(t)}{T_2(t)} + \frac{\Delta}{4}, T_2(t) = l)$$

Using the fact that $S_2(t)/T_2(t)$ is an unbiased estimator of $\frac{1}{l}\sum_{m=1}^{l} X_{2,m}$

$$\implies \mathbb{P}(A(t), T_2(t) \geq L, \theta_2(t) > \mu_2 + \frac{\Delta}{2}) \leq \sum_{t=1}^{T} \mathbb{P}(\theta_2(t) > \frac{1}{l}\sum_{m=1}^{l} X_{2,m} + \frac{\Delta}{4}, T_2(t) = l) \tag{19.3}$$

recall that,

$$\theta_2(t)|_{S_2(t),F_2(t)} \sim Beta(1 + S_2(t), 1 + F_2(t))$$

Conditioning over S we have RHS of 19.3 ,

$$= \sum_{l=L}^{T}\sum_{S=1}^{l} \mathbb{P}(\sum_{m=1}^{l} X_{2,m} = S) * \mathbb{P}((\theta_2(t) > \frac{1}{l}\sum_{m=1}^{l} X_{2,m} + \frac{\Delta}{4}, T_2(t) = l| \sum_{m=1}^{l} X_{2,m} = S)) \tag{19.4}$$

Using, $\mathbb{P}(A,B|C) = \mathbb{P}(B|C)\mathbb{P}(A|B,C) \implies \mathbb{P}(A,B|C) \leq \mathbb{P}(A|B,C)$ for the last term

$$\leq \sum_{l=L}^{T}\sum_{S=1}^{l} \mathbb{P}(\sum_{m=1}^{l} X_{2,m} = S) * \mathbb{P}(\theta_2(t) > \frac{S}{l} + \frac{\Delta}{4}|T_2(t) = l, \sum_{m=1}^{l} X_{2,m} = S)$$
$$= \sum_{l=L}^{T} \mathbb{E}_{S\sim Bin(l,\mu_2)}[\mathbb{P}(Beta(1+S, 1+l-S) > \frac{S}{l} + \frac{\Delta}{4})]$$

Neat fact about the beta distribution
$F_{Beta(a,b)}(y) = 1 - F_{Bin(a+b-1,y)}(a-1)$
$\mathbb{P}(Beta(1+S, 1+l-S) > \frac{S}{l} + \frac{\Delta}{4}) \equiv F_{Bin(l+1,y)}(S) \leq F_{Bin(l,y)}$, (for notational convenience)
$= \mathbb{P}[\sum_{i=1}^{l} U_i \leq S]$ where $U_i \sim Ber(y)$
$= \mathbb{P}(\frac{1}{l}\sum_{i=1}^{l} U_i - y \leq \frac{S}{l} - y)$
using Azuma Hoeffding inequality

$$\leq exp(\frac{-2l\Delta^2}{16}) \leq exp(\frac{-2L\Delta^2}{16}) = \frac{1}{T^3}$$

$\therefore$ Second term $\leq \frac{1}{T^2}$
Putting it together

$$\mathbb{P}(E(t)^C) \leq \frac{2}{T^2}$$
$$\implies \mathbb{P}(E(t)) \leq 1 - \frac{2}{T^2}$$

Lemma 2: Deals with bounding the average.

$$\mathbb{E}(W(j, S_{1,j}, y) \cap T) \leq \begin{cases} 1 + \frac{2}{1-y} + \frac{\mu_1}{\Delta'} e^{-Dj} & j < \frac{y}{D} log(R) \\ 1 + \frac{R^y}{1-y} e^{-Dj} + \frac{\mu_1}{\Delta'} e^{-Dj} & \frac{y}{D} log(R) \leq j < \frac{4log(T)}{\Delta'^2} \\ \frac{16}{T} & j \geq \frac{4log(T)}{\Delta'^2} \end{cases} \tag{19.5}$$

Proof : Exercise.
Lemma 3: For all non-negative integers j , $S \leq j$, and for all $y \in [0,1]$,

$$\mathbb{E}(W(j, S_{1,j}, y) \cap T | S_{1,j}) = \frac{1}{F_{Bin(j+1,y)}(S)} - 1$$

Proof: Exercise
Two sources of randomness
1. Randomness in $S_{1,j}$
2. Randomness in W so we can write the above expression as

$$\mathbb{E}_{S_{1,j} \sim Bin(j,\mu_1)}(\mathbb{E}(W(j, S_{1,j}, y) \cap T | S_{1,j}))$$

which is equivalent to,

$$\mathbb{E}_{S_{1,j}\sim Bin(j,\mu_1)}(\mathbb{E}(Geo(1-F_{Beta(1+S,1+j-S)}(y))))-1$$
$$=\mathbb{E}_{S_{1,j}\sim Bin(j,\mu_1)}[\frac{1}{F_{Bin(j+1,y)}(S)}]-1$$

[Proof left as an exercise]
Reference:[1]

19.1.4 Regret Analysis for 2 Arm case

Using equation 19.1 with Lemma 1, 2, 3 we get

$$\mathbb{E}(T_2(T)) \leq L+\sum_{j=0}^{T}\mathbb{E}(W(j,S_{1,j},y)\cap T)+T\sum_{j=0}^{T}\mathbb{E}(\sum_{t=v_j+1}^{v_{j+1}-1}\mathbb{1}(\{\theta_2(t)\geq y, j\geq J_0\})) \tag{19.6}$$

We use Lemma 1 to bound the last term, lemma 2 to bound the second term, finally get,

$$\mathbb{E}(T_2(T)) \leq \frac{40*log(T)}{\Delta^2}+\frac{48}{\Delta^4}+18$$

Detailed proof is left as an Exercise. Expected regret can be bounded as,

$$\mathbb{E}(\mathbb{R}_T)=\mathbb{E}(\Delta*T_2(T))=\frac{40*log(T)}{\Delta}+\frac{48}{\Delta^3}+18*\Delta$$

For N arms similar argument holds.

19.1.5 Wrapping Up

Thompson Sampling performance.

1. [1] $[Agarwal-Goyal'2011]$: Rewards $\in$ [0,1],
   for N=2, expected regret $O(\frac{log(T)}{\Delta})$
   for general N Regret=$O((\sum_{i\neq i^*}\frac{1}{\Delta^2})^2 log(T))$
   Not better than UCB but very promising approach.

2. [2] $[Kaufmann-Korda-Munos'12]$ Bernaulli Bandits
   $\forall\varepsilon>0$
   Expected Regret at time T $\leq (1+\varepsilon)\sum_{i\neq i^*}\frac{\Delta_i(log(T)+loglog(T))}{D(\mu_i||\mu_{i^*})}+const(\varepsilon,\mu_1,\mu_2,...,\mu_N)$
   Asymptotically optimal with respect to time.
   $Note$: Asymptotic regret scaling of $(\sum_{i\neq i^*}\frac{\Delta_i}{D(\mu_i||mu_{i^*})})log(T)$ is known to be Optimal. [Lai-Robbins-1985]

3. [1] $[Agarwal - Goyal'2011]$: Bernaulli Rewards
   $\forall \varepsilon > 0$
   $\mathbb{E}(\mathbb{R}_T) \leq (1+\varepsilon) \sum_{i \neq i^*} \frac{\Delta_i}{D(\mu_i || \mu^*)} * log(T) + O(\frac{N}{\varepsilon^2})$
   Extended to much more general families.

4. [3] $[kaufmann - et - al'2013]$: Continuous reward distribution
   Reward Dist $\in$ 1-dimensional exponential family [Beta, Gaussian, gamma,Pareto...]
   $\forall \varepsilon > 0$:
   $\mathbb{E}(\mathbb{R}_T) \leq (\frac{1+\varepsilon}{1-\varepsilon}) \frac{(\mu^* - \mu_i)}{D(\theta_i || \theta^*)} * log(T) + const(\varepsilon)$

References

[1] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011.

[2] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, pages 199–213, 2012.

[3] Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1448–1456, 2013.