IE613: Online Machine Learning **Jan-Apr 2018**

Lecture 23: Thompson Sampling and UCB-V

Lecturer: M. K. Hanawal *Scribes: Nishant Mani Tripathi*

23.1 Thompson Sampling for the Multi-armed Bandit Problem

[1] **The multi-armed bandit problem :** In MAB we are given a slot machine with N arms, at each time step t = 1, 2, 3, . . ., one of the N arms must be chosen to be played. Each arm i, when played, yields a random real-valued reward according to some fixed distribution which is not known with support in [0, 1]. The random reward obtained from playing an arm repeatedly are i.i.d. and independent of the plays of the other arms. The reward is observed immediately after playing the arm. So, the different learning algorithm chooses an arm based on some decision parameter using the information available for t-1 round.

Thompson Sampling, is a natural Bayesian algorithm. The basic idea is to choose an arm to play according to its probability of being the best arm. Thompson Sampling algorithm has experimentally been shown to be close to optimal.Thompson Sampling algorithm achieves logarithmic expected regret for the stochastic multi-armed bandit problem.

The Thompson Sampling algorithm initially assumes arm i to have prior Beta(1, 1) on μ_i. At time t, having observed $S_i(t)$ successes (reward = 1) and $F_i(t)$ failures (reward = 0) in $k_i(t) = S_i(t) + F_i(t)$ plays of arm i, the algorithm updates the distribution on i as Beta$(S_i(t) + 1, F_i(t) + 1)$. The algorithm then samples from these posterior distributions of the μ_i, and plays an arm according to the probability of its mean being the largest.

Thompson Sampling algorithm for the Bernoulli bandit problem(when the rewards are either 0 or 1). The algorithm for Bernoulli bandits maintains Bayesian priors on the Bernoulli means μ_i

Algorithm 1 Thompson Sampling for Bernoulli bandits

Input : K, T
Initialize : $i \in [K], S_i = 0, F_i = 0$
For each: t=1,2,...,T
$\forall i \in [K]$, sample $\theta_i(t)$ from $Beta(S_i + 1, F_i + 1)$ distribution play arm $I_t = argmax_i \theta_i(t)$ and observe $\tilde{r_t}$

```
if r_t = 1 then
    S_{I_t} = S_{I_t} + 1
else
    F_{I_t} = F_{I_t} + 1
end if
```

Assume that First arm is the only optimal arm. The assumption of unique optimal arm is without loss of generality, since adding more arms with $\mu_i = \mu^*$ can only decrease the expected regret. Therefore

Theorem 23.1 *If $K = 2$, the expected regret of Thomson sampling is*

$$\bar{R_T} = O\left(\frac{\ln T}{\Delta} + \frac{1}{\Delta^3}\right)$$

in time T, $\Delta = \mu_1 - \mu_2$

Theorem 23.2 *If $K > 2$, the expected regret of Thomson sampling is*

$$\bar{R_T} \leq O\left(\left(\sum_{i \neq i^*} \frac{1}{\Delta_i^2}\right) \log T\right)$$

in time T, $\Delta = \mu_1 - \mu_i$

Modify Thomson sampling such that after observing the reward $\tilde{r_t} \in [0, 1]$ at time t, it performs a Bernoulli trial with success probability $\tilde{r_t}$. Let random variable r_t denote the outcome of this Bernoulli trial, and let $\{S_i(t), F_i(t)\}$ denote the number of successes and failures of Bernoulli trials till time t. Rest algorithm is the same as for Bernoulli bandits.

Algorithm 2 Thompson Sampling for General stochastic Multi Arm Bandits

Input : K, T
Initialize : $i \in [K], S_i = 0, F_i = 0$
For each: t=1,2,...,T
$\forall i \in [K]$, sample $\theta_i(t)$ from $Beta(S_i + 1, F_i + 1)$ distribution play arm $I_t = argmax_i \theta_i(t)$ and observe $\tilde{r_t}$
Perform Bernoulli trail with success probability $\tilde{r_t}$ and observe $r_t \in \{0, 1\}$

```
if r̃_t = 1 then
    S_{I_t} = S_{I_t} + 1
else
    F_{I_t} = F_{I_t} + 1
end if
```

Let f_i denote the (unknown) pdf of reward distribution for arm i. Then, on playing arm i,

$$\mathbb{E}[r_t | \tilde{r_t}] = \tilde{r_t}$$

$$\mathbb{E}[r_t] = \int_0^1 \tilde{r_t} f_{I_t}(\tilde{r_t}) d\tilde{r_t} = \mu_{I_t}$$

23.2 Varients of UCB

$$argmax_i \left(B_{i,t} = \hat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{\alpha \log T}{T_i(t-1)}} \right)$$

23.2.1 UCB-V

UCB-V policy:

$$argmax_i \left(B_{i,t} = \mu_{i,T_i(t-1)} + \sqrt{\frac{2V_{i,T_i(t-1)}\varepsilon_{T_i(t-1),t}}{T_i(t-1)}} + \frac{3c\varepsilon_{T_i(t-1),t}}{T_i(t-1)} \right)$$

At time t, play an arm maximizing $B_{i,t}$.
[2] The UCB-V policy uses the function ε, to facilitate exploration. Let us summarize the main ideas underlying the algorithm. Till the time a arm is not chosen its bound is infinite. Hence, at the start the algorithm selects all the arms at least once one after the other. After this initial phase the arms will be tried multiple times. The more an arm k has been tested, the closer the bound gets to the sample-mean, and hence, by the law of large numbers, to the expected reward μ_k. So the procedure will hopefully tend to draw arms having the largest expected rewards with an increasing frequency.
Since the rewards generated are stochastic it is possible that during the first draws the optimal arm always gives low rewards. This will make the sample mean of that arm smaller than that of the others. Hence an algorithm that only uses sample means might not choose the optimal arm any more. UCB policies generally prevent this situation by using upper confidence bounds on the mean rewards. The confidence level with which these bounds hold determine the amount of exploration of the policy and ultimately the performance of the algorithm.

Assumption
Let $K > 2$ and let $\upsilon_1, ..., \upsilon_K$ be distributions over the real with support $[0, b]$. For $1 \leq k \leq K$, let $X_{k,t} \sim \upsilon_k$ be an i.i.d. sequence of random variables specifying the rewards for arm k. Assume that the rewards of different arms are independent, i.e., for any $t \leq 1$, the vectors $(X_{1,1}, ..., X_{1,t}), ..., (X_{K,1}, ..., X_{K,t})$ are independent. The decision maker does not know the distributions of the arms, but knows b.

$$argmax_i \left(B_{i,t} = \mu_{i,T_i(t-1)} + \sqrt{\frac{2V_{i,T_i(t-1)}\varepsilon_{T_i(t-1),t}}{T_i(t-1)}} + \frac{3c\varepsilon_{T_i(t-1),t}}{T_i(t-1)} \right)$$

where,
$\hat{\mu}_{i,s} = \frac{\sum_{t=1}^{s} X_{i,t}}{s}$ $\hat{V}_{i,s} = \frac{\sum_{t=1}^{s}(X_{i,t}-\mu_{i,s})^2}{s}$
$\varepsilon_{s,t}$ is increasing in t
let $\varepsilon_{s,t} = \beta \log T$

so we get,

$$argmax_i \left(\hat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{2\beta \log t V_{i,T_i(t-1)}}{T_i(t-1)}} + \frac{3c\beta \log t}{T_i(t-1)} \right)$$

$$\beta > 1, c > 0$$

Theorem 23.3 *When $c = 1$ and $\beta = 1.2$, the expected regret of UCB-V is*

$$\bar{R_T} \leq \sum_{k \neq k^*} \left(\frac{\sigma_k^2}{\Delta_k} + 2 \right) \log T$$

References

[1] Shipra Agrawal and Navin Goyal, 2012, Analysis of Thompson Sampling for the Multi-armed Bandit Problem, JMLR: Workshop and Conference Proceedings vol 23 (2012) 39.139.26

[2] Audibert, J.Y., Munos, R. and Szepesvri, C., 2009. Explorationexploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science, 410(19), pp.1876-1902.