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

# Lecture 21: UCB

*Lecturer: M. K. Hanawal* *Scribes: Vinay Chourasiya*

**Disclaimer**: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

## 21.1 Introduction:

In previous class, we discussed about stochastic MAB setting.In this setting,each arm $i \in [K]$ is associated with an (unknown) probability distribution $\nu_i$ and reward of each arm is i.i.d from $\nu_i$. we stated some methods for stochastic MAB setting such as e-greedy, softmax, Bayesian exploration, optimistic exploration. we also stated the regret bound for UCB(1) algorithm i.e. $O(\frac{K \log T}{\Delta})$.

$$\tilde{R_T} = \sum_{i=1}^{K} \mathbb{E}\left[N_{i,(T)}\right] \Delta_i$$

In this lecture we consistence with previous lecture notation.In this lecture will discuss about the arm pulling policy of the UCB(1) algorithm.In the calculation of pseudo regret we need $\mathbb{E}\left[N_{i,(T)}\right]$ where $i \neq i^*$,so we will find out the expected number of times sub-optimal arm pulled by the algorithm.

## 21.2 UCB

$I_t$ = arm pulled by the UCB algorithm at round $t$

$$I_t = \arg\max_i \left\{ \hat{u}_{i,N_{i,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i,(t-1)}}} \right\} \tag{21.1}$$

$N_{i,(t-1)}$ = random variable that count how many times arm $i$ pulled in $(t-1)$ round

**Theorem:** *The expected number of pull of sub-optimal arm $i$ after $T$ round, $\Delta_i = \mu^* - \mu_i$*

$$\mathbb{E}\left[N_{i,(T)}\right] \leq \frac{4\alpha \log T}{\Delta_i^2} + \frac{\pi^2}{3} + 1$$

$$\begin{aligned} \tilde{R_T} &= \sum_{k=1}^{K} \mathbb{E}\left[N_{i,(T)}\right] \Delta_i \\ &\leq \sum_{i \neq i^*} \frac{4\alpha \log T}{\Delta_i} + k(\frac{\pi^2}{3} + 1) \end{aligned}$$

$$\leq \sum_{i \neq i^*} \frac{4\alpha \log T}{\Delta_i} + C \tag{21.2}$$

From (21.2), *We can say that, in total round $T$, Number of times optimal arm pull( $i^*$ )*

$$T - (k-1)\left(\frac{4\alpha \log T}{\Delta_i^2} + (\frac{\pi^2}{3} + 1)\right)$$

**Proof :** In round $t$ , $i \neq i^*$ ( define $I_t = i$) is played $(t > k)$ if arm $i$ ( sub-optimal arm) chosen then equation (21.3) hold

$$\hat{u}_{i,N_{i,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i,(t-1)}}} \geq \hat{u}_{i^*,N_{i^*,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i^*,(t-1)}}} \tag{21.3}$$

and also if arm $i$ (i.e. sub-optimal) pulled then at least one of the following inequality hold.

$$\hat{u}_{i^*,N_{i^*,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i^*,(t-1)}}} < \mu^* \tag{21.4}$$

$$\hat{u}_{i,N_{i,(t-1)}} - \sqrt{\frac{\alpha \log t}{N_{i,(t-1)}}} > \mu_i \tag{21.5}$$

$$N_{i,(t-1)} \leq \frac{4\alpha \log T}{\Delta_i^2} \tag{21.6}$$

if equation (21.4), ( 21.5) and ( 21.6) does not hold, then

$$\hat{u}_{i^*,N_{i^*,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i^*,(t-1)}}} \geq \mu^*$$

$$\hat{u}_{i,N_{i,(t-1)}} - \sqrt{\frac{\alpha \log t}{N_{i,(t-1)}}} \leq \mu_i$$

$$N_{i,(t-1)} > \frac{4\alpha \log T}{\Delta_i^2}$$

Let arm $i$ (sub-optimal one) pulled and given above inequality (21.4),(21.5) and (21.6) not hold then, we get,

$$\hat{u}_{i,N_{i,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i,(t-1)}}} \leq \mu_i + 2\sqrt{\frac{\alpha \log t}{N_{i,(t-1)}}} \qquad since\ (21.5) false$$

$$\left\{2\sqrt{\frac{\alpha \log t}{N_{i,(t-1)}}} < \Delta_i,\ since\ (21.6)\ is\ false\ and\ \mu_i = \mu_i^* - \Delta_i\right\}$$

$$\leq \mu_i^* - \Delta_i + \Delta_i$$

$$\leq \hat{u}_{i^*,N_{i^*,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i^*,(t-1)}}} \qquad From\ 21.4$$

Which is contradicting (21.3), hence if sub-optimal $i$ pulled then atleast one of (21.4),(21.5) and (21.6) hold.

Define: $u = \lceil \frac{4\alpha \log T}{\Delta_i^2} \rceil$ and $u < T$

Now, We calculate Expected number of time sub-optimal arm pulled i.e.

$$\begin{aligned}\mathbb{E}\left[N_{i,(T)}\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{(I_t=i)}\right] \\ &= \mathbb{E}\left[N_{i,(u)}\right] + \mathbb{E}\left[\sum_{t=u+1}^{T} \mathbb{1}_{(I_t=i)}\right]\end{aligned}$$

Put the value of $u$

$$= \lceil \frac{4\alpha \log T}{\Delta_i^2} \rceil + \mathbb{E}\left[\sum_{t=u+1}^{T} \mathbb{1}_{(I_t=i \ and \ (21.6) \ is \ false)}\right]$$

$$= \lceil \frac{4\alpha \log T}{\Delta_i^2} \rceil + \mathbb{E}\left[\sum_{t=u+1}^{T} \mathbb{1}_{(I_t=i \ and \ (21.4) \ or \ 21.5 \ hold)}\right]$$

$$= \lceil \frac{4\alpha \log T}{\Delta_i^2} \rceil + \mathbb{E}\left[\sum_{t=u+1}^{T} \mathbb{1}_{(I_t=i \ and \ (21.4) \ hold)}\right] + \mathbb{E}\left[\sum_{t=u+1}^{T} \mathbb{1}_{(I_t=i \ and \ (21.5) \ hold)}\right]$$

$$= \lceil \frac{4\alpha \log T}{\Delta_i^2} \rceil + \sum_{t=u+1}^{T} P\left\{I_t = i \ and \ (21.4) \ hold\right\} + \sum_{t=u+1}^{T} P\left\{I_t = i \ and \ (21.5) \ hold\right\} \tag{21.7}$$

here,

$$P\left\{I_t = i \ and \ (21.4) \ hold\right\} \leq P\left\{\hat{u}_{i^*,N_{i^*,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i^*,(t-1)}}} < \mu^*\right\}$$

$$\left[we \ know \ that \ P\left\{\hat{\mu} - \mu > \epsilon\right\} \leq \exp(-2n\epsilon^2)\right]$$

$$\begin{aligned}&\leq P\left\{\exists s \in \{1, 2..t\} \ \hat{\mu}_{i^*,s} + \sqrt{\frac{\alpha \log t}{s}} < \mu^*\right\} \\ &\leq P\left\{\hat{\mu}_{i^*,s} + \sqrt{\frac{\alpha \log t}{s}} < \mu^*\right\} \\ &\leq \sum_{s=1}^{t} P\left\{\hat{\mu}_{i^*,s} + \sqrt{\frac{\alpha \log t}{s}} < \mu^*\right\} \\ &\leq \sum_{s=1}^{t} P\left\{\hat{\mu}_{i^*,s} - \mu^* < \sqrt{\frac{\alpha \log t}{s}}\right\}\end{aligned}$$

using Hoeffding's Inequality

$$\leq \sum_{s=1}^{t} \exp(-2s(\frac{\alpha \log t}{s}))$$

$$= \sum_{s=1}^{t} t^{-2\alpha}$$

$$\leq \frac{1}{t^{2\alpha-1}}$$

similarly we can show,

$$P\left\{I_t = i \textit{ and } (21.5) \textit{ hold}\right\} \leq P\left\{\hat{u}_{i^*,N_{i^*,(t-1)}} + \sqrt{\frac{\alpha \log t}{N_{i^*,(t-1)}}} < \mu^*\right\}$$

$$\leq \frac{1}{t^{2\alpha-1}}$$

Back to equation (21.7)

$$\mathbb{E}\left[N_{i,(T)}\right] = \lceil \frac{4\alpha \log T}{\Delta_i^2} \rceil + 2\sum_{t=u+1}^{T} \frac{1}{t^{2\alpha-1}}$$

$$\left[2\alpha - 1 > 2, \quad since\ \alpha > \frac{3}{2}\right]$$

$$\leq \lceil \frac{4\alpha \log T}{\Delta_i^2} \rceil + 2\sum_{t=u+1}^{T} \frac{1}{t^2}$$

$$\leq \frac{4\alpha \log T}{\Delta_i^2} + \frac{\pi^2}{3} + 1 \qquad \left\{\sum_{t=u+1}^{T} \frac{1}{t^2} \to \frac{\pi^2}{6}\right\}$$

Hence Proved.

## Case Analysis :

Case-1: let $\mu_1, \mu_2, \mu_3$ mean of 1,2,3 arm respectively and $\Delta = \arg\min_i \Delta_i$

Case-2: and $\bar{\mu_1}, \bar{\mu_2}, \bar{\mu_3}$ mean of three different arm and $\bar{\Delta} = \arg\min_i \Delta_i$

Qus. : *if $\bar{\Delta} < \Delta$ then, in which case regret will be more ?*

We know that,

$$\tilde{R_T} = \sum_{i=1}^{K} \mathbb{E}\left[N_i(T)\right] \Delta_i$$

$$\leq min\left\{T\Delta, O(\frac{K \log T}{\Delta})\right\}$$

Regret depend upon value of T, if $T$ is small ,then $T\Delta$ dominate.
if $T >> \frac{1}{\Delta}$, than $O(\frac{K \log T}{\Delta})$ dominate.

## 21.3 References:

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2):235256, 2002.