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

Lecture 20: Stochastic Multi-Armed Bandits

Lecturer: Stochastic Multi-Armed Bandits0 *Scribes: Sudhindra Katre*

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.*

20.1 Stochastic Multi-armed Bandits

The stochastic bandit problem:

Known parameters: number of arms 'K'
Unknown parameters: K probability distributions $\nu_1, \nu_2, ..., \nu_k$ on [0,1].
For each round t=1,2,...,T:
1)The learner/forecaster chooses $I_t \epsilon 1, 2, ..., K$
2) Given I_t, the environment draws reward $X_{I_t}(t) \sim \nu_{I_t}$
The learner/forecaster wants the highest expected reward.

Let μ_i be the expectation of the $i^t h$ arm.

$$\mu_i = \mathbb{E}[\nu_i] \qquad\qquad i = 1, 2, ..., K \tag{1}$$

The total reward till round T is $= \sum_{t=1}^{T} X_{I_t}(t)$
If the learner had always pulled the best arm, i^*,

$$i^* = \underset{1}{\mathrm{argmax}} \mu_i$$

Reward in that case would be $= \max_{1} \sum_{t=1}^{T} X_i(t)$ Using these, the expected regret can be defined as

$$R_T = \mathbb{E}[\max_{1} \sum_{t=1}^{T} X_i(t)] - \mathbb{E}[\sum_{t=1}^{T} X_{I_t}(t)]$$

Pseudo regret is defined as

$$\tilde{R}_T = \max_{1} \mathbb{E}[\sum_{t=1}^{T} X_i(t)] - \mathbb{E}[\sum_{t=1}^{T} X_{I_t}(t)]$$

In both definitions, the expectation is taken with respect to the random draw of both rewards and forecaster's actions. Substituting (1),

$$\tilde{R}_T = T\mu^* - \sum_{t=1}^{T} \mu_{I_t}$$

Introducing the term $N_i(T)$ in the equation above, where $N_i(T)$ is the number of pulls of arm i till round T.

$$\tilde{R}_T = \mathbb{E}[\sum_{i=1}^{K} N_i(T)]\mu^* - \mathbb{E}[\sum_{i=1}^{K} N_i(T)\mu_i]$$

$$\implies \tilde{R}_T = \sum_{i=1}^{K} \mathbb{E}[N_i(T)](\mu^* - \mu_i)$$
$$\implies \tilde{R}_T = \sum_{i=1}^{K} \mathbb{E}[N_i(T)]\Delta_i \tag{2}$$

where $\Delta_i = \mu^* - \mu_i$ are the 'gaps' between the mean of best arm and mean of other arms.

Note that the expectation here is over realisation of I_t.

20.2 Algorithms for Stochastic MAB

Defining $P_i(t+1)$ as the probability for drawing arm i in $(t+1)^{th}$ round. This is updated in t^{th} round. Defining mean estimate of i^{th} arm till round t as

$$\hat{\mu}_i(t) = \frac{\sum_{s=1}^{t} X_{I_s}(s)\mathbb{1}_{\{I_s=i\}}}{\sum_{s=1}^{t} \mathbb{1}_{\{I_s=i\}}}$$

$\epsilon-$greedy algorithm
Update rule:

$$P_i(t+1) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{K} & if\ i = \text{argmax}\ \hat{\mu}(t) \\ \frac{\epsilon}{K} & otherwise \end{cases}$$

Remarks:
If $\epsilon = constant$ for all rounds, it will give linear regret (not sub-linear) for all values of ϵ.If ϵ is varied with rounds, we can achieve sub-linear regret.
Softmax algorithm

$$P_i(t+1) = \frac{e^{\frac{\hat{\mu}_i(t)}{\tau}}}{\sum_{j=1}^{K} e^{\frac{\hat{\mu}_j(t)}{\tau}}}$$

Here τ is the Boltzman temperature parameter.
Remarks
As $T \to \infty$, we get a uniform distribution.
As $T \to 0$, the term with highest mean will dominate and will go towards greedy algorithm.
Bayesian exploration
Start with a prior and pull posterior distribution (Thompson sampling).
Optimistic exploration algorithms
Optimism in the face of uncertainty (OFU) based on upper confidence bounds (UCB).

20.3 UCB

Change in notation

$$\hat{\mu}_{i,s} = \frac{\sum_{n=1}^{s} X_{i,n}}{s}$$

$X_{i,n}$ is n^{th} sample of $i^{th} arm$.

Algorithm 1 UCB

Parameters : $\alpha > \frac{3}{2}, K = number of arms$
Let P_1 be the uniform distribution over $\{1, 2, .., K\}$
Initialize: Pull each arm once.
For each round t =K+1,K+2,...,T :
Draw arm I_t, given by

$$I_t = \underset{1}{\text{argmax}}[\hat{\mu}_{i,N_i(t-1)} + \sqrt{\frac{\alpha \log t}{N_i(t-1)}}]$$

Remarks:
The first term in the expression for I_t is the exploitation term, while the second term is exploration term. Note that exploration is done even for high $N_i(t-1)$, due to the $\log t$ term in the numerator of the second term.

Theorem 20.1. *UCB Theorem*
For the UCB algorithm as given above, following relations hold:
1. $\mathbb{E}[N_i(T)] \leq \frac{4\alpha \log T}{\Delta_i^2} + (\frac{\pi^2}{3} + 1)$
2. $\tilde{R}_T \leq \sum_{i \neq i^*} \frac{4\alpha \log T}{\Delta_i} + (\frac{\pi^2}{3} + 1)K$

$\therefore \tilde{R}_T \leq \frac{4\alpha \log T}{\Delta} + (\frac{\pi^2}{3} + 1)K \qquad$ *where* $\Delta = \min_{i \neq i^*} \Delta_i$

Remarks:
1. Δ is called the sub-optimality gap. If Δ is small, regret is high.
2. Since the regret bound depends on Δ, this is a problem dependent bound.
3. From the theorem, we have $\mathbb{E}[N_i(T)] \leq \frac{4\alpha \log T}{\Delta_i^2} + (\frac{\pi^2}{3} + 1)$

   From (2), we have $\tilde{R}_T = \sum_{i=1}^{K} \mathbb{E}[N_i(T)]\Delta_i$

Using this and the expression for $\mathbb{E}[N_i(T)]$ from the inequality in the theorem,

$\tilde{R}_T = \sum_{i=1}^{K} \mathbb{E}[N_i(T)]\Delta_i$
$\tilde{R}_T = \sum_{i \neq i^*} \mathbb{E}[N_i(T)]\Delta_i$
$\tilde{R}_T \leq \sum_{i \neq i^*} [\frac{4\alpha \log T}{\Delta_i^2} + (\frac{\pi^2}{3} + 1)]\Delta_i$
$\tilde{R}_T \leq \sum_{i \neq i^*} [\frac{4\alpha \log T}{\Delta_i} + (\frac{\pi^2}{3} + 1)\Delta_i]$
$\tilde{R}_T \leq \sum_{i \neq i^*} [\frac{4\alpha \log T}{\Delta} + (\frac{\pi^2}{3} + 1)K]$

(Using the inequality $\sum_{i \neq i^*} \Delta_i \leq K$ (since $\Delta_i \leq 1 \; \forall \; i$) and Δ is as defined before).
Thus, it is enough to prove the first statement of the proof, as the other statements follow from it.

Getting rid of Δ_i

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

Rewriting for using Cauchy Schwartz inequality,

$$\tilde{R}_T = \sum_{i=1}^{K} \sqrt{\mathbb{E}[N_i(T)]}\sqrt{\mathbb{E}[N_i(T)]\Delta_i^2}$$

$$\tilde{R}_T \leq \sqrt{\sum_{i=1}^{K} \mathbb{E}[N_i(T)] \;\; \sum_{i=1}^{K} \mathbb{E}[N_i(T)]\Delta_i^2}$$

Substituting $\sum_{i=1}^{K} \mathbb{E}[N_i(T)] = T$ and the expression for $\mathbb{E}[N_i(T)]$ from the first statement of the theorem,

$$\tilde{R}_T \leq \sqrt{T \;\; \sum_{i=1}^{K} [\tfrac{4\alpha \log T}{\Delta_i^2} + (\tfrac{\pi^2}{3} + 1)]\Delta_i^2}$$

$$\therefore \quad \tilde{R}_T \leq \sqrt{T \;\; \sum_{i=1}^{K} [4\alpha \log T + (\tfrac{\pi^2}{3} + 1)\Delta_i^2]}$$

$$\therefore \quad \tilde{R}_T \leq \sqrt{T[4K\alpha \log T + K(\tfrac{\pi^2}{3} + 1)}$$

$$\therefore \quad \tilde{R}_T \leq \sqrt{TK[4\alpha \log T + (\tfrac{\pi^2}{3} + 1)} \qquad \text{(since } \textstyle\sum_{i=1}^{K} \Delta_i^2 \leq K\text{)}$$

Thus we now have a problem independent bound. Proof of UCB theorem will be done in the next lecture.