# Week 8 – Lecture 36

## Divergence

Let $p$ & $q$ be any distribution, then divergence between them is defined as:

$$d(P,q) = \sum_{i=1}^{N} P_i \log \frac{P_i}{q_i} \geq 0$$

Proof: of $d(P,q) \geq 0$

$$d(P,q) = -\sum_{i=1}^{N} P_i \log \frac{q_i}{P_i}$$

Apply Jensen's inequality

---

## KL-UCB Algorithm

Input: No. of arms = $K$, constant $c$

Horizon = $T$

Initialize: Play each arm once

For $t = K+1, K+2, \ldots T$

$$I_t \leftarrow \underset{a \in [K]}{\arg\max} \; \max \left\{ q \in \Theta : N_a(t-1)\, d\left(\frac{S_a(t-1)}{N_a(t-1)}, q\right) \leq \log t + c \log(\log t) \right\}$$

index for each arm

$r \leftarrow$ observed reward from $I_t$

$N_{I_t}(t) \leftarrow N_{I_t}(t-1) + 1$ : No. of pulls of arms

$S_{I_t}(t) \leftarrow S_{I_t}(t-1) + r$ : Cumulative reward

$\frac{S_a(t-1)}{N_a(t-1)}$ & $q$ are Bernouli distributions

Divergence for Bernouli :

$p, (1-p)$

$q, (1-q)$

$$d(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{(1-p)}{(1-q)}$$

If $q \in [p, 1]$, $\therefore$ all $q$s that are larger than $p$

then $d(p,q) = 0$ for $q = p$

and increasing for all $q \in [p, 1]$

$\therefore$ If $q \in \left[ \frac{S_a(t-1)}{N_a(t-1)}, 1 \right]$

Then the divergence of the index will increase till it reaches a point where the inequality is violated.

Thus at the point the inequality is violated, that would be max $q$, & and that $q$ will be the upper confidence bound.

---

<u>Theorem</u>: Consider a K-armed stochastic bandit with support in $[0, 1]$.

Let $\varepsilon > 0$ & set $c = 3$.

For any $T$, the number of times KL-UCB chooses a sub-optimal arm $K \, (\neq K^*)$ is bounded as

$$\mathbb{E}\left[N_k(T)\right] \leq \frac{\log T \; (1+\varepsilon)}{d(\mu_K, \mu_{K^*})} + C_1 \log(\log T) + \frac{C_2(\varepsilon)}{T^{\beta(\varepsilon)}}$$

← log(log T will be very small thus can ignore

largest term compare with UCB log T term

T is already in denominator ∴ ignore

where $C_1$ is constant

$C_2$ and $\beta$ are positive functions of $\varepsilon$

---

<u>Lemma</u>: <u>Pinsker Inequality</u> [Proof is found in information

... performance ...
Theoretic analysis
or in the book ]

$\forall\ p, q \in [0, 1]$

$$d(p,q) \geq 2(p-q)^2$$

$$\therefore\ d(\mu_k, \mu_{k^*}) \geq 2(\mu_k - \mu_{k^*})^2$$
$$= 2\Delta_k^2$$

---

$$\mathbb{E}[T_k(n)] \leq \frac{6 \log n}{\Delta_k^2} + \pi^2/3 + 1$$

for UCB

$$\mathbb{E}[N_k(T)] \leq \frac{\log T}{d(\mu_k, \mu_{k^*})}(1+\varepsilon) + C_1 \log(\log T) + \frac{C_2(\varepsilon)}{T^{\beta(\varepsilon)}}$$

for KL-UCB

$$\leq \frac{\log T}{2\Delta_k^2}(1+\varepsilon) + C_1 \log(\log T) + \frac{C_2(\varepsilon)}{T^{\beta(\varepsilon)}}$$

Thus we have a better bound on the expected number of pulls $\therefore$ would have a better regret bound.

---