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

Lecture 22: Kullback Leibler-UCB (KL-UCB)

Lecturer: M. K. Hanawal *Scribes: Karan Patel*

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.*

22.1 Recap:

- In previous class mainly we focused on how to compute the expected number of pulls of sub-optimal arm after T rounds.

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

- Also we can conclude that $\left(T - (k-1)\left(\frac{4\alpha \log T}{\Delta_i^2} + (\frac{\pi^2}{3} + 1)\right)\right)$ will be the number of times the optimal arm was pulled after T rounds.
- Also we have seen that pseudo-regret calculation depends on $\mathbb{E}\left[N_{i,(T)}\right]$ and it can be calculated as follows:

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

22.2 UCB

Using previous class results we know that Pseudo regret is given by:

$$\tilde{R}_T \leq \sum_{i \neq i^*} \left(\frac{4\alpha \log T}{\triangle_i^2} + \frac{\pi^2}{3} + 1\right) \Delta_i$$

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

Let $B_{i,t} = \hat{\mu}_{i,N_i(t-1)} + \sqrt{\frac{\alpha \log t}{N_i(t-1)}}$

Now by using Hoeffding's inequality, $Pr(\mathbb{E}[x] \geq \epsilon) \leq 2e^{-2n\epsilon^2}$, probability of $B_{i,t}$ can be written as follows:

$$Pr \left\{ \hat{\mu}_{i,N_i(t-1)} + \sqrt{\frac{\alpha \log t}{N_i(t-1)}} \gtreqless \mu_i \right\} \approx \exp\left(-2N_i(t-1)\frac{\alpha \log t}{N_i(t-1)}\right)$$

$$= \frac{1}{t^{2\alpha}}$$

From above simplification of UCB equation we can see that $\log t$ term in UCB ensuring that confidence term $(\frac{1}{t^{2\alpha}})$ is not constant but depends on 't'.

22.3 Optimistic Algorithm

22.3.1 Kullback Leibler-UCB (KL-UCB):

Algorithm : KL-UCB

Input: T (Horizons), K (Number of arms)
Initialize: Play each arm once and observe rewards.
for t = K+1,K+2,...,T In above equation 'd' represents divergence. Let here

$$B_{i,t} = max \left\{ q \in [0,1], d(\hat{\mu}_{i,N_i(t-1)}, q) \leq \frac{\log t + c\log(\log t)}{N_i(t-1)} \right\}$$

Where, 'd' represents the KL-divergence which is given as follows:

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

We also know that,

$$d(p,q) = \begin{cases} \geq 0, & p \neq q \\ 0, & p = q \end{cases}$$

Further if we fix p and look divergence as function of other variable then for a given p, $d(p,q)$ is strictly convex in q and increasing in the interval $q \in (p,1)$ as shown in following figure;

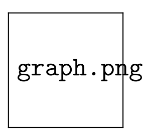

Figure 22.1: Function $d(p,q)$ for a given p

Fixed p in above figure is : $\frac{\log t + \log(\log t)}{N_i(t-1)}$.

Theorem 22.1 *Consider a bandit problem with K arms and independent rewards bounded in [0,1]. Let* $\epsilon > 0$ *and c=3. Let* i^* *be the arm with maximal expected reward* μ_i^* *and* i *be a sub-optimal arm (i.e.* $i \neq i^*$ *or* $\Delta^* \neq 0$*).For any positive integer T, the number of times algorithm KL-UCB chooses arm i is upper-bounded by:*

$$\mathbb{E}[N_i(T)] \leq \frac{\log T}{d(\mu_i, \mu_i^*)}(1+q) + C_1\log(\log T) + \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}} \qquad \forall \epsilon > 0$$

where $C_1, C_2(\epsilon)$ *and* $\beta(\epsilon)$ *are positive functions of* ϵ *(¿ 0), and T is time horizon.*

Divide both side by $\log T$*, we get*

$$\frac{\mathbb{E}[N_i(T)]}{\log T} \leq \frac{1}{d(\mu_i, \mu_i^*)}(1+q) + \frac{C_1\log(\log T)}{\log T} + \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}(\log T)}$$

Taking $\limsup_{T\to\infty}$ *both side, we get*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\log T} \leq \frac{1}{d(\mu_i, \mu_i^*)}$$

NOTE: Pinsker's inequality, $d(\mu_i, \mu_i^*) > 2(\mu_i^* - \mu_i)^2$.

$$\boxed{\textbf{For UCB: } \limsup_{T\to\infty} \tfrac{\mathbb{E}[N_i(T)]}{\log T} \leq \tfrac{8\alpha}{2\Delta_i^2}}$$

And thus KL-UCB has strictly better upper bound (asymptotically) than UCB, while it has the same range of application.

The regret of the KL-UCB algorithm stisfies :

$$\limsup_{T\to\infty} \frac{\tilde{R}_T}{\log T} \leq \sum_{i:i\neq i^*} \frac{\mu_{i^*} - \mu_i}{d(\mu_i, \mu_i^*)} = \sum_{i:i\neq i^*} \frac{\Delta_i}{d(\mu_i, \mu_i^*)}$$

$$\boxed{\textbf{For UCB: } \limsup_{T\to\infty} \tfrac{\mathbb{E}[N_i(T)]}{\log T} \leq \tfrac{8\alpha \sum_{i:i\neq i^*}(\mu_i^* - \mu_i)}{2\Delta_i^2} \leq 8\alpha \sum_{i\neq i^*} \tfrac{\Delta_i}{\Delta_i^2}}$$

22.3.2 Lower Bound:

Irrespective of algorithm we use, the following holds:

$$\liminf_{T\to\infty} \frac{\tilde{R_T}}{\log T} \leq \sum_{i\neq i^*} \frac{\Delta_i}{d(\mu_i, \mu_i^*)}$$

Here, $\frac{\tilde{R_T}}{\log T}$:= Regret is normalized by $\log T$.

i.e. The KL-UCB algorithm thus appears to be (asymptotically) optimal.

Define: KL-UCB relies on the following upper-confidence bound for μ_i:

$$B_{i,t} := u_i(t) = max \left\{ q > \hat{\mu}_{i,N_i(t-1)} : d(\hat{\mu}_i(t), q) \leq \frac{\log t + 3\log(\log t)}{N_i(t-1)} \right\}$$

For $x, y \in [0,1]$, define $d^+(x,y) = d(x,y)\ \mathbb{1}x < y$. Without loss of generality, let $i^* = 1$. The expectation of $N_i(T)$ is upper-bounded by using the following:

$$\mathbb{E}[N_i(T)] = \mathbb{E}[\sum_{t=1} \mathbb{1}\{I_t = i\}]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{I_t = i, \ \ \mu_1 > u_1(t), \ \ \mu_1 \leq u_1(t)\}\right]$$

(here, μ_1 = true value and $u_1(t)$= index).

After decomposed above can be written as:

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mu_1 > u_1(t)\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{I_t = i, \ \ \mu_1 \leq u_1(t)\}\right] \tag{22.1}$$

Lemma 22.1

$$\sum_{t=1}^{T} \mathbb{1}\{I_t = i, \;\; \mu_1 \leq u_1(t)\} \leq \sum_{s=1}^{T} \mathbb{1}\{sd^+(\hat{\mu}_{i,s}, \mu_1) < \log t + 3\log(\log t)\}$$

Proof Let we pull some arm i in t_{th} round (i.e. $I_t = i$) such that $\mu_1 < u_1(t)$ which together implies that $u_i(t) \geq u_1(t) > \mu_1$ hence,

$$d^+(\hat{\mu}_i(t), \mu_1) \leq d(\hat{\mu}_i(t), \mu_i(t) = \frac{\log t + 3\log(\log t)}{N_i(t)}$$

From above equation we can write

$$\sum_{t=1}^{T} \mathbb{1}\{I_t = i, \; \mu_1 \leq u_1(t)\} \leq \sum_{t=1}^{T} \mathbb{1}\{I_t = i, \; N_i(t)d^+(\hat{\mu}_i(t), \mu_1) \leq \log t + 3\log(\log t)\}$$

$$= \sum_{t=1}^{T}\sum_{s=1}^{t} \mathbb{1}\{N_t(i) = s, \; I_t = i, \; sd^+(\hat{\mu}_{i,s}, \mu_1) \leq \log t + 3\log(\log t)\}$$

$$\leq \sum_{t=1}^{T}\sum_{s=1}^{t} \mathbb{1}\{N_t(i) = s, \; I_t = i, \; \}\mathbb{1}\{sd^+(\hat{\mu}_{i,s}, \mu_1) \leq \log t + 3\log(\log t)\}$$

$$= \sum_{s=1}^{T} \mathbb{1}\{sd^+(\hat{\mu}_{i,s}, \mu_1) \leq \log t + 3\log(\log t)\} \sum_{t=s}^{T} \mathbb{1}\{N_t(i) = s, \; I_t = i\}$$

as, for every $s \in \{1, 2, ..., T\}$, $\sum_{t=s}^{T} \mathbb{1}\{N_t(i) = s, \; I_t = i\} = 1$

$$= \textstyle\sum_{s=1}^{T} \mathbb{1}\{sd^+(\hat{\mu}_{i,s}, \mu_1) \leq \log t + 3\log(\log t)\}$$

By using above lemma, equation (22.1) can be written as follows:

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mu_1 > u_1(t)\}\right] + \mathbb{E}\left[\sum_{s=1}^{T} \mathbb{1}\{sd^+(\hat{\mu}_{i,s}, \mu_1) < \log t + 3\log(\log t)\}\right]$$

$$\leq \sum_{t=1}^{T} \mathbb{P}(\mu_1 > u_1(t)) + \mathbb{E}\left[\sum_{s=1}^{T} \mathbb{1}\{sd^+(\hat{\mu}_{i,s}, \mu_1) < \log t + 3\log(\log t)\}\right]$$

The first summand is upper-bounded as follows:

$$\mathbb{P}(\mu_1 > u_1(t)) \leq \frac{e\lceil \log(t)^3 + 3\log(\log(t))\rceil}{t\log(t)^3}$$

For the second summand let

$$K_T = \lfloor \frac{1+\epsilon}{d^+(\mu_i, \mu_1)}(\log T + 3\log(\log T))\rfloor \tag{22.2}$$

then

$$\begin{aligned}
&\sum_{s=1}^{T} \mathbb{P}\left(sd^{+}(\hat{\mu}_{i,s}, \mu_1) < \log T + 3\log(\log T)\}\right) \\
&\leq K_T + \sum_{s=K_T+1}^{\infty} \mathbb{P}\left(sd^{+}(\hat{\mu}_{i,s}, \mu_1) < \log t + 3\log(\log t)\right) \\
&\leq K_T + \sum_{s=K_T+1}^{\infty} \mathbb{P}\left(K_T d^{+}(\hat{\mu}_{i,s}, \mu_1) < \log t + 3\log(\log t)\right)
\end{aligned}$$

by using equation(22.2) above inequality can be written as:

$$\leq K_T + \sum_{s=K_T+1}^{\infty} \mathbb{P}\left(d^{+}(\hat{\mu}_{i,s}, \mu_1) < \frac{d(\mu_i, \mu_1)}{1+\epsilon}\right) \tag{22.3}$$

Lemma 22.2 *For each $\epsilon > 0$, there exist $C_2(\epsilon) > 0$ and $\beta(\epsilon) > 0$ such that*

$$\sum_{s=K_T+1}^{\infty} \mathbb{P}\left(d^{+}(\hat{\mu}_{i,s}, \mu_1) < \frac{d(\mu_i, \mu_1)}{1+\epsilon}\right) \leq \frac{C_2(\epsilon)}{T^{\beta()\epsilon}}$$

Proof If $d^{+}(\hat{\mu}_{i,s}, \mu_1) < \frac{d(\mu_i,\mu_1)}{1+\epsilon}$, then $\hat{\mu}_{i,s} > r(\epsilon)$ where $r(\epsilon) \in (\mu_i, \mu_1)$ such that $d\left(r(\epsilon), \mu_1\right) = \frac{d(\mu_i,\mu_1)}{1+\epsilon}$. Hence,

$$\begin{aligned}
\mathbb{P}\left(d^{+}(\hat{\mu}_{i,s}, \mu_1) < \frac{d(\mu_i, \mu_1)}{1+\epsilon}\right) &\leq \mathbb{P}\left(d(\hat{\mu}_{i,s}, \mu_i) > d\left(r(\epsilon), \mu_i\right), \mu_{i,s} > \mu_i\right) \\
&\leq \mathbb{P}\left(\mu_{i,s} > r(\epsilon)\right) \leq exp\left(-sd(r(\epsilon), \mu_i)\right)
\end{aligned}$$

and

$$\sum_{s=K_T+1}^{\infty} \mathbb{P}\left(d^{+}(\hat{\mu}_{i,s}, \mu_1) < \frac{d(\mu_i, \mu_1)}{1+\epsilon}\right) \leq \frac{exp\left(-d(r(\epsilon), \mu_i) K_T\right)}{1 - exp\left(-d(r(\epsilon), \mu_i)\right)} \leq \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}}$$

$$\begin{aligned}
with \quad C_2(\epsilon) &= \frac{1}{1 - exp\left(-d(r(\epsilon), \mu_i)\right)} \approx O(\epsilon^{-2}) \\
\beta(\epsilon) &= \frac{(1+\epsilon) d(r(\epsilon), \mu_i)}{d(\mu_i, \mu_1)} \approx O(\epsilon^{2})
\end{aligned}$$

By using result of above lemma, equation(22.3) can be written as:

$$\leq \frac{1+\epsilon}{d^{+}(\mu_i, \mu_1)} (\log T + 3\log(\log T)) + \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}}$$

References

[1] Garivier, Aurlien, and Olivier Capp. "The KL-UCB algorithm for bounded stochastic bandits and beyond." In Proceedings of the 24th annual Conference On Learning Theory, pp. 359-376. 2011.