KL-UCB for Bernoulli r.vs.

Source: Chap 10, Bandit Algorithms, L. & S (textbook)

A. Garivier & O. Cappé, The KL-UCB algorithm for bounded stochastic bandits and beyond, COLT 2011

So far, general sub-Gaussian noise, and upper bounds. This class: More refined algorithm and analysis for Bernoulli r.vs,

i.e. $X_j \sim \text{Bernoulli}(\mu_j)$, $\mu_j \in [0,1]$, $j = 1, 2, \dots, k$

Why should we expect better?

$$\text{Var}(X_j) = \mu_j(1-\mu_j) \leq \frac{1}{4} \quad \forall \mu_j \in [0,1].$$

In UCB, we (implicitly) used that variance ≤ 1 to design the exploration bonus. While that assumption is satisfied here, estimating the mean μ_j also gives us a refined estimate

on the variance of this particular environment instance, i.e., of $\sigma_j = \sqrt{\mu_j(1-\mu_j)}$.

Thus, we can further adapt the exploration bonus based on samples.

Relative Entropy / KL Divergence for Bernoulli r.vs.

Given two finite pmfs $p(x_i), q(x_i)$, $i=1,2,\cdots,L$ with $q(x_i) > 0 \;\forall\; i=1,2,\cdots,L$,

$$KL(p(\cdot)\|q(\cdot)) = \sum_{i=1}^{L} p(x_i) \ln \frac{p(x_i)}{q(x_i)}$$

For the special case of Bernoulli r.vs, with parameters $p, q \in [0,1]$,

$$KL(p(\cdot), q(\cdot)) = d(p,q)$$

$$= p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right).$$

$d(0,q) = \ln\left(\frac{1}{1-q}\right)$, $d(1,q) = \ln(1/q)$ $d(0,0) = 0$

$d(p,0) = \infty$, $d(1,1) = 0$, $d(p,1) = \infty$

Properties of (Bernoulli) KL-Divergence.

(a) $d(\cdot, q)$ and $d(p, \cdot)$ are convex, with unique minima at q and p respectively.

(b) (Pinsker's Inequality): $d(p,q) \geq 2(p-q)^2$

(c) $p \leq q - \varepsilon \leq q$,

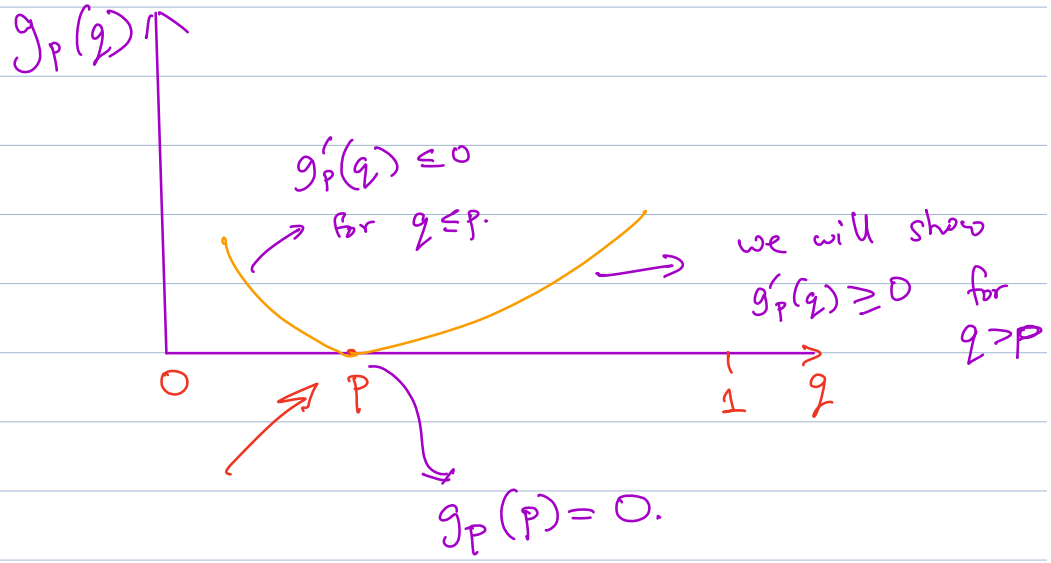
$$d(p, q-\varepsilon) \leq d(p,q) - 2\varepsilon^2.$$

Proof: (a) $\rightarrow$ check by taking second derivative and testing for positivity. $d(p,p) = d(q,q) = 0$, function strictly positive elsewhere.

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

Pf: fix any $p \in (0,1)$, and let

$$g_p(q) = p\ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right) - 2(p-q)^2$$



$g_p(p) = 0$ (substitute).

$$\frac{d}{dq} g_p(q) = -\frac{p}{q} + \frac{(1-p)}{(1-q)} + 4(p-q)$$

$$= \frac{pq - p + q - pq}{q(1-q)} - 4(q-p)$$

$$= (q-p)\left(\underbrace{\frac{1}{q(1-q)}}_{\ge 4} - 4 \right)$$

≥ 0.

$$\therefore \quad \frac{d}{dq} g_p(q) \begin{cases} \geq 0 & \text{for } q > p \\ \leq 0 & \text{for } q < p \\ = 0 & \text{for } q = p \end{cases}$$

$$\Rightarrow \quad g_p(q) \geq 0 \qquad \blacksquare$$

(Thanks to A. Mazumdar's notes on Applied Info Theory @ U. Mass, Feb. 2016).

ⓒ $h(p) = d(p, q-\varepsilon) - d(p,q)$

linear in p (with positive slope)

$$= p \ln \frac{q}{q-\varepsilon} + (1-p) \ln \frac{1-q}{1-q+\varepsilon}$$

$$= \ln\left(\frac{1-q}{1-q+\varepsilon}\right) + p \ln\left(\underbrace{\frac{q}{q-\varepsilon}}_{>1} \cdot \underbrace{\frac{1-q+\varepsilon}{1-q}}_{>1}\right)$$

$$\underbrace{\qquad\qquad}_{>0}$$

Also, $p \leq q-\varepsilon \leq q$:

$$\therefore \quad h(p) \leq h(q-\varepsilon) = -d(q-\varepsilon, q).$$

$$\overset{\text{Pinsker's ineq.}}{\leq} -2\varepsilon^2$$

Chernoff Bound for iid Bernoulli r.v.s:

$\{X_i, i=1,2,\ldots,X_n\}$ iid $\sim$ Bernoulli(μ).

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$P(\hat{\mu}_n \geq \mu+\varepsilon) \leq e^{-n\,d(\mu+\varepsilon, \mu)}$$

$$P(\hat{\mu}_n \leq \mu-\varepsilon) \leq e^{-n\,d(\mu-\varepsilon, \mu)}$$

Pf: HW Problem!

(Write the rate function, and simplify).

Corollary: $P\left(d(\hat{\mu}_n, \mu) \geq a, \hat{\mu}_n \leq \mu\right) \leq e^{-na}$

$P\left(d(\hat{\mu}_n, \mu) \geq a, \hat{\mu}_n \geq \mu\right) \leq e^{-na}$

"inverting" the $d(\cdot, \mu)$ function

Pf: $d(x, \mu)$ is monotone decreasing over $x \in [0, \mu]$.

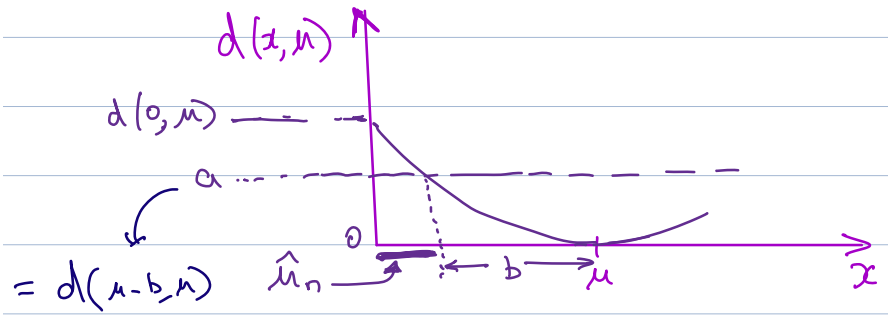
$\therefore$ Choose any $a \in [0, d(0, \mu)]$:

$$\left\{ d(\hat{\mu}_n, \mu) \geq a, \hat{\mu}_n \leq \mu \right\}$$

$$= \left\{ \hat{\mu}_n \leq \mu - b, \hat{\mu}_n \leq \mu \right\}$$

$$= \left\{ \hat{\mu}_n \leq \mu - b \right\}$$

where b is the unique soln to $d(\mu - b, \mu) = a$.



Recall: $P(\hat{\mu}_n - \mu \leq -\varepsilon) \leq e^{-n d(\mu-\varepsilon, \mu)}$

$\therefore P(d(\hat{\mu}_n, \mu) \geq a, \hat{\mu}_n \leq \mu)$

$= P(\hat{\mu}_n - \mu \leq -b)$

$\leq e^{-n d(\mu-b, \mu)}$

$= e^{-na}$

(QED)

The **KL-UCB Algorithm**

Setting: K arms, unstructured,

$P_{a_i} \sim \text{Bernoulli}(\mu_i)$, $\mu_i \in [0,1]$.

$\mu_1 > \mu_2 \geq \mu_3$ ----

$f(t) = 1 + t \ln^2(t)$

(a) Play each arm once

(b) For $t \geq k+1$,

$$A_t = \arg\max_j \max \left\{ \tilde{\mu}_j \in [0,1] : d\left(\hat{\mu}_j(t-1), \tilde{\mu}_j\right) \leq \frac{\ln f(t)}{T_j(t-1)} \right\}$$

$$\hat{\mu}_j(t-1) = \frac{1}{T_j(t-1)} \sum_{s=1}^{t-1} X_s \, \mathbb{1}_{\{A_s = j\}}$$

Theorem: Setting as above. (10.6 in book)

$$R_n(\pi, \nu) \leq \sum_{\{j : \Delta_j > 0\}} \inf_{\substack{\varepsilon_1, \varepsilon_2 > 0 \\ \varepsilon_1 + \varepsilon_2 \in (0, \Delta_j)}} \Delta_j \left(\frac{\ln f(n)}{d(\mu_j + \varepsilon_1, \mu_1 - \varepsilon_2)} + \frac{1}{2\varepsilon_1^2} + \frac{2}{\varepsilon_2^2} \right)$$

Also, $$\overline{\lim_{n \to \infty}} \frac{R_n}{\ln(n)} \leq \sum_{\{j : \Delta_j > 0\}} \frac{\Delta_j}{d(\mu_j, \mu_1)}$$

Intuition: Let $\tilde{\mu}_j^* = \underset{\tilde{\mu}_j \in [\hat{\mu}_j, 1]}{\arg\max}\ d(\hat{\mu}_{j(t-1)}, \tilde{\mu}_j)$

$$\leq \frac{\ln f(t)}{T_j(t-1)}$$

Then, the interpretation is the following: Samples from arm j could have been generated by $\{0,1\}$ r.v.s $\{X_1, X_2, \ldots\}$ iid, s.t. $X_r \sim \text{Bernoulli}(\tilde{\mu}_j^*)$

$\tilde{\mu}_j^*$ is a plausible "upper confidence" mean, with the confidence defined through the KL divergence, and the level chosen at $\left(\frac{\ln f(t)}{T_j(t-1)}\right)$.

Thus, instead of considering a std-dev based intuition in UCB $\left(\text{where } \sqrt{\frac{2\ln f(t)}{T_j(t-1)}}\right)$ was the exploration bonus, we are using the KL divergence to provide such a bonus.

Note: We will see later that this is asymptotically tight.

Discussion: Benefit wrt UCB occurs when $\mu_j \neq \frac{1}{2}$.

We have $X_i \sim \text{Bernulli}(\mu_i)$ and μ_i unknown

$\Rightarrow$ $X_j \sim \frac{1}{2}$ − subGaussian (worst case bound on std- dev).

UCB regret with $\sigma = \frac{1}{2}$ gives:

$$\overline{\lim_{n\to\infty}} \frac{R_n}{\ln(n)} \leq \sum_{j:\Delta_j>0} \frac{1}{2\Delta_j}$$

KL- UCB:

$$d(\mu_j, \mu_1) = \mu_j \ln \frac{\mu_j}{\mu_1} + (1-\mu_j) \ln \frac{(1-\mu_j)}{(1-\mu_1)}$$

Suppose $\mu_1 - \mu_j = \Delta_j \approx$ small

Taylor's series expansion to the second term:

$$d(\mu_j, \mu_1) = -\mu_j \ln \frac{\mu_1}{\mu_j} - (1-\mu_j) \ln \frac{(1-\mu_1)}{(1-\mu_j)}.$$

$$= -\mu_j \ln\left(1 + \frac{\Delta_j}{\mu_j}\right) - (1-\mu_j) \ln\left(\frac{1-\mu_j-\Delta_j}{1-\mu_j}\right)$$

Taylor's Thm (up to quadratic term)

$$\approx -\mu_j \left[\frac{\Delta_j}{\mu_j} + \frac{\Delta_j^2}{2\mu_j^2}\right] - (1-\mu_j)\left[\frac{-\Delta_j}{1-\mu_j} + \frac{\Delta_j^2}{2(1-\mu_j)^2}\right]$$

$$= \frac{\Delta_j^2}{2\mu_j} - \frac{\Delta_j^2}{2(1-\mu_j)} = \frac{\Delta_j^2}{2\mu_j(1-\mu_j)}$$

$\therefore$

$$d(\mu_j, \mu_1) = \frac{\Delta_j^2}{2\mu_j(1-\mu_j)} + o(\Delta_j^2)$$

KL UCB

$\therefore$ We expect: $\lim_{n\to\infty} \frac{R_n}{\ln(n)} \lesssim \sum_{j:\Delta_j>0} \frac{2\mu_j(1-\mu_j)}{\Delta_j}$

If $\mu_j = \frac{1}{2}$: $\frac{2\mu_j(1-\mu_j)}{\Delta_j} = \frac{1}{2\Delta_j}$ = constant with UCB

For all other cases, KL-UCB improves the constant. We will later see that this matches the lower bound.

Next, let $X_r \sim \text{Bernoulli}(\mu)$, $r = 1, 2, \dots, n$ (iid), and $\varepsilon > 0$, $\mu \in [0,1]$. Define:

$$\tau = \min\left\{ t \geq 1 : \max_{1 \leq r \leq n} \left(\underline{d}\left(\hat{\mu}_r, \mu - \varepsilon\right) - \frac{\ln f(t)}{r} \right) \leq 0 \right\}$$

(random variable; time; samples)

where $\underline{d}(p,q) = d(p,q)\, \mathbb{1}_{\{p \leq q\}}$.

Interpretation: Suppose $X_r \equiv X_{1r}$, i.e., samples from arm 1 (the best arm). Then consider any time $t \geq \tau$.

We have for $t = \tau$,

$$\max_{1 \leq r \leq n} \left(\underline{d}\left(\hat{\mu}_{1r}, \mu_1 - \varepsilon\right) - \frac{\ln f(t)}{r} \right) \leq 0$$

Further, note $\ln f(t)$ is increasing in t.

$\Rightarrow \forall \; t \geq \tau,$

$$\max_{1 \leq r \leq n} \left(\frac{d\left(\hat{\mu}_{1r}, \mu_1 - \varepsilon\right) - \frac{\ln f(t)}{r}}{} \right) \leq 0.$$

i.e., $$\max_{1 \leq r \leq n} \left(d\left(\hat{\mu}_{1r}, \mu_1 - \varepsilon\right) \chi_{\{\hat{\mu}_{1r} \leq \mu_1 - \varepsilon\}} - \frac{\ln f(t)}{r} \right) \leq 0$$

Thus at any time t s.t. $n \geq t \geq \tau$

$$d\left(\hat{\mu}_{1 T_1(t-1)}, \mu_1 - \varepsilon\right) \chi_{\{\hat{\mu}_{1 T_1(t-1)} \leq \mu_1 - \varepsilon\}}$$

$T_1(t-1)$: number of samples of arm 1 until time t

$$\leq \frac{\ln f(t)}{T_1(t-1)}$$

$\therefore$ Either $\hat{\mu}_{1 T_1(t-1)} > \mu_1 - \varepsilon$ (case 1: $\chi_{\{\cdot\}} = 0$)

OR (case 2) $d\left(\hat{\mu}_{1 T_1(t-1)}, \mu_1 - \varepsilon\right) \leq \frac{\ln f(t)}{T_1(t-1)}$

Recall KL-UCB: we pick the arm with largest $\tilde{\mu}_j^*$ defined by:

$$\tilde{\mu}_j^* = \operatorname{argmax}\left\{\tilde{\mu}\in[0,1]: d\left(\hat{\mu}_{j\,T_j(t-1)}, \tilde{\mu}\right) \le \frac{\ln f(t)}{T_j(t-1)}\right\}$$

Case 1: $\hat{\mu}_{1\,T_1(t-1)} > \mu_1 - \varepsilon \implies \tilde{\mu}_1^* > \mu_1 - \varepsilon$

Case 2: $d\left(\hat{\mu}_{1\,T_1(t-1)}, \mu_1 - \varepsilon\right) \le \frac{\ln f(t)}{T_1(t-1)}$

$$\Rightarrow \quad \tilde{\mu}_1^* \ge \mu_1 - \varepsilon$$

$\Rightarrow$ The index of the played arm A_t at time $t \ge \tau$ satisfies:

$$\tilde{\mu}_{A_t}^* \ge \tilde{\mu}_1^* \ge \mu_1 - \varepsilon.$$

Suppose $A_t = j$. Then, by KL-UCB

$$d\left(\hat{\mu}_{j\,T_j(t-1)}, \tilde{\mu}_j^*\right) \le \frac{\ln f(t)}{T_j(t-1)}$$

for some $\tilde{\mu}_j^* \in [\mu_1 - \varepsilon, 1]$.

Lemma (10.7 in textbook) $\{X_1, \cdots, X_n\} \sim \text{Bernoulli}(\mu)$, iid; $\varepsilon > 0$.

$$\tau = \min\left\{ t \geq 1 : \max_{1 \leq r \leq n} \frac{d}{r}\left(\hat{\mu}_r, \mu - \varepsilon\right) - \frac{\ln f(t)}{r} \leq 0 \right\}$$

Then $$E[\tau] \leq \frac{2}{\varepsilon^2}$$

Proof:

$$P(\tau > t) \leq P\left(\bigcup_{r=1}^{n} \left\{ \frac{d}{r}\left(\hat{\mu}_r, \mu - \varepsilon\right) > \frac{\ln f(t)}{r} \right\} \right)$$

$$\leq \sum_{r=1}^{n} P\left(\frac{d}{r}\left(\hat{\mu}_r, \mu - \varepsilon\right) > \frac{\ln f(t)}{r} \right)$$

$$= \sum_{r=1}^{n} P\left(\left\{ d\left(\hat{\mu}_r, \mu - \varepsilon\right) > \frac{\ln f(t)}{r} \right\} \cap \left\{ \hat{\mu}_r < \mu - \varepsilon \right\} \right)$$

$$\leq \sum_{r=1}^{n} P\left(d\left(\hat{\mu}_r, \mu\right) > \frac{\ln f(t)}{r} + 2\varepsilon^2 , \ \hat{\mu}_r < \mu \right)$$

↳ KL divergence (iii) ; Pinsker's Lemma.

$$\leq \sum_{r=1}^{n} e^{-r\left(2\varepsilon^2 + \frac{\ln f(t)}{r}\right)}$$

inverting KL concentration (Corollary 10.4 in text)

$$\leq \frac{1}{f(t)} \sum_{r=1}^{n} e^{-2r\varepsilon^2} \leq \frac{1}{2f(t)\varepsilon^2}$$

$$\therefore E[\tau] \leq \int_0^\infty P(\tau > t)\, dt \leq \frac{1}{2\varepsilon^2} \int_1^\infty \frac{dt}{f(t)}$$

$E[\tau]$: non-negative rv

$$\leq \frac{2}{\varepsilon^2}$$ ∎

Recall that the above construction ensures that for $t \geq \tau$, we have that whenever arm j is played, $d\left(\hat{\mu}_{j T_j(t-1)}, \tilde{\mu}_j^*\right) \leq \frac{\ln f(t)}{T_j(t-1)}$

with $\tilde{\mu}_j^* \geq \mu_1 - \varepsilon$.

We need to next bound the number of times this occurs.

The discussion below proves a slightly more general version of Lemma 10.8, because I think that this is needed in the main theorem.

Caveat: I could not convince myself that Lemma 10.8 in book suffice for one of the steps (I will highlight as (**) later). But, it might .... let me know if it does!

Lemma 10.8* Let $\{X_1, X_2, \ldots, X_n\}$ be iid r.vs., $X_i \sim \text{Bernoulli}(\mu)$, $\mu \in [0,1]$.

* Let $Y_r = g_r(X_1, \ldots, X_r)$ be non-negative r.vs, with $Y_r \in [0,1]$., $r = 1, 2, \ldots, n$,

i.e., Y_r is measurable wrt $\sigma\{X_1, \ldots, X_r\}$.

Fix $\Delta > 0$, $a > 0$, and

$$K = \sum_{r=1}^{n} \chi\left\{ d(\hat{\mu}_r, \mu + \Delta + Y_r) \leq \frac{a}{r} \right\}.$$

* additional term not in text

Then $E[\tau] \le \inf_{\varepsilon \in (0,\Delta)} \left(\frac{1}{2\varepsilon^2} + \frac{a}{d(\mu+\varepsilon, \mu+\Delta)} \right)$

Proof: Let $\varepsilon \in (0, \Delta)$

$$E[K] = \sum_{r=1}^{n} P\left(d(\hat{\mu}_r, \mu+\Delta+Y_r) \le \frac{a}{r} \right)$$

Let $A_r = \{ \hat{\mu}_r > \mu + \varepsilon \}$, $r = 1, 2, \dots, n$

$$= \sum_{r=1}^{n} P\left(A_r \cap \left\{ d(\hat{\mu}_r, \mu+\Delta+Y_r) \le \frac{a}{r} \right\} \right)$$

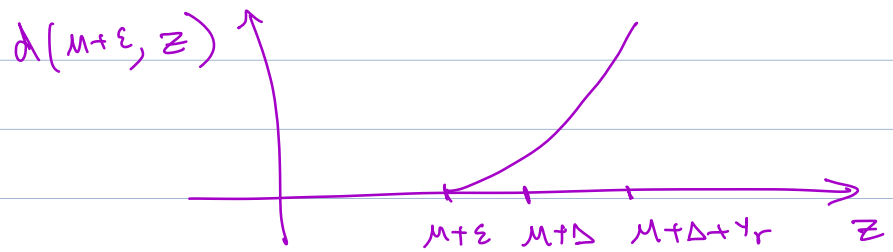
$$+ \sum_{r=1}^{n} P\left(A_r^c \cap \left\{ d(\hat{\mu}_r, \mu+\Delta+Y_r) \le \frac{a}{r} \right\} \right)$$

$\hat{\mu}_r = x \le \mu+\varepsilon$

$d(x, \mu+\Delta+Y_r)$ decreasing in $x \in [0, \mu+\Delta]$ $\quad > \varepsilon$

$$\le \sum_{r=1}^{n} P(A_r) + \sum_{r=1}^{n} P\left(d(\mu+\varepsilon, \underbrace{\mu+\Delta+Y_r}_{\ge \mu+\Delta}) \le \frac{a}{r} \right)$$

$d(\mu+\varepsilon, \mu+\Delta+(\cdot)) \le \frac{a}{r} \implies d(\mu+\varepsilon, \mu+\Delta) \le a/r$



$$\leq \sum_{r=1}^{n} P\left(\hat{\mu}_r > \mu+\varepsilon\right) + \sum_{r=1}^{n} P\left(d(\mu+\varepsilon, \mu+\Delta) \leq \frac{a}{r}\right)$$

either 1 or 0 for each r as there is no r s.t. $d(\mu+\varepsilon, \mu+\Delta) \leq \frac{a}{r}$

$$\leq \sum_{r=1}^{n} e^{-r\, d(\mu+\varepsilon, \mu)} + \left(\frac{a}{d(\mu+\varepsilon, \mu+\Delta)}\right).$$

$$\leq \frac{1}{d(\mu+\varepsilon, \mu)}$$

Pinsker's ineq.

$$\leq \frac{1}{2\varepsilon^2} + \frac{a}{d(\mu+\varepsilon, \mu+\Delta)}.$$

$$\Rightarrow \mathbb{E}[K] \leq \inf_{\varepsilon \in (0,\Delta)} \left(\frac{1}{2\varepsilon^2} + \frac{a}{d(\mu+\varepsilon, \mu+\Delta)}\right)$$

∎

The main result: Setting as before with k Bernoulli(μ_j) arms.

$$R_n \le \sum_{\{j:\Delta_j>0\}} \inf_{\substack{\varepsilon_1,\varepsilon_2>0 \\ 0<\varepsilon_1+\varepsilon_2<\Delta_j}} \Delta_j \left(\frac{\ln f(n)}{d(\mu_j+\varepsilon_1, \mu_1-\varepsilon_2)} + \frac{1}{2\varepsilon_1^2} + \frac{2}{\varepsilon_2^2} \right)$$

Proof: We will bound $E[T_j(n)]$, and use regret decomposition to complete the proof.

μ_1 best arm mean

$\mu_j = \mu_1 - \Delta_j$, a sub-optimal arm.

$$E[T_j(n)] = E\left[\sum_{t=1}^{n} \chi_{\{A_t=j\}}\right]$$

$$= E\left[\sum_{t=1}^{\tau} \chi_{\{A_t=j\}}\right] + E\left[\sum_{t=\tau+1}^{n} \chi_{\{A_t=j\}}\right]$$

$$\le E[\tau] + E\left[\sum_{t=\tau+1}^{n} \chi_{\{A_t=j\}}\right]$$

(**)

Recall for $t \geq \tau+1$ that $A_t = j$

$$\Rightarrow \underbrace{d\left(\hat{\mu}_{j\, T_j(t-1)}, \tilde{\mu}_j^*\right)}_{B_{j,t}} \leq \frac{\ln f(t)}{T_j(t-1)}$$

for some $\tilde{\mu}_j^* \geq (\mu_1 - \varepsilon)$.

$$\therefore E[T_j(n)] \leq E[\tau] + E\left[\sum_{t=\tau+1}^{n} \chi_{\{A_t = j \,\cap\, B_{j,t}\}}\right]$$

$$\leq E[\tau] + E\left[\sum_{t=1}^{n} \chi_{B_{j,t}}\right]$$

Using the same trick as in the UCB proof (page 12 in Notes: 5 - UCB fixed horizon) to switch from time steps to number of samples, we can replace $T_j(t-1) \longrightarrow r$, and sum over $r = 1, 2, \cdots, n$.

(**)

Now, use the Lemma* 10.8 to bound the second term.

$$E[T_j(n)] \leq E[c] + E[k].$$

Substitute the two bounds to get the result.