The UCB Algorithm

Upper Confidence Bound

Source: Chap 7, Bandit Algorithms. (textbook)

So far:

ETC: Number of samples per arm predetermined.
Δ, n known

ETC + Doubling Trick: Δ known, horizon arbitrary.

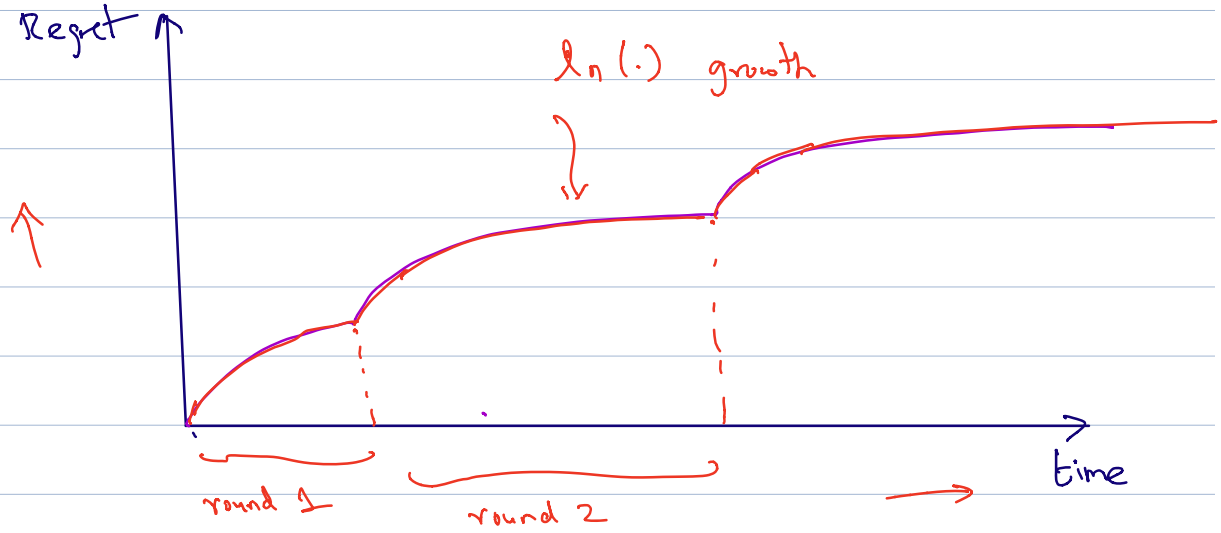
ε-Greedy: Samples used for estimation (aka exploration samples) seperated from time slots used for regret optimization (aka exploitation samples)
Δ known, n is arbitrary.

Elimination algo'. In each phase, number of samples per remaining arm is predetermined
Δ unknown, n is fixed

Elimination + Doubling trick: Δ unknown, n arbitrary.

All algos above: Samples for exploration seperate from samples from exploitation.

The "problem" with doubling trick is that regret is not "smooth"



UCB Algorithm: Δ unknown, n arbitrary, regret growth is smooth

Key difference from prev algorithms: Number of samples for exploration for each arm is dynamic, and no seperation between exploration and exploitation, i.e., all samples are used to improve estimates.

Technical Challenge: Empirical estimates of arms

are constructed using random number of samples, so usual sub-Gaussian concentration do not hold.

* Need anytime concentrations:

Let $\{X_i, i=1,2,\dots\}$ sequence of iid 1-SubGaussian r.vs., and τ a r.v.

Then $$P\left(\left|\sum_{i=1}^{\tau} X_i\right| > \sqrt{2\tau\left(2\log\tau + \log(1/\delta)\right)}\right) \leq \frac{\pi^2}{3}\delta.$$

(τ: r.v)

Compare to usual 1-SubGaussian bound:

$$P\left(\sum_{i=1}^{n} X_i > \sqrt{2n\log(1/\delta)}\right) \leq \delta$$

(n: deterministic)

UCB relies on Principle of Optimism: Use the "most reasonable" optimistic estimate of each arm, and optimize your decision at each time-step using this optimistic estimate.

UCB algorithm studied in two steps.

First: $\{\Delta_i\}$ unknown, n = time-horizon KNOWN

we will develop the basic algo with this assumption, and then modify the algo (next class) to eliminate this requirement.

Setting: Unstructured environment, k arms, $P_{a_i} \sim (\mu_i + 1\text{-subGaussian})$ reward.

$$U_j(t-1, \delta) = \underbrace{\hat{\mu}_j(t-1)}_{\text{empirical mean}} + \underbrace{\sqrt{\frac{2\ln(1/\delta)}{T_j(t-1)}}}_{\text{exploration bonus}} \dashrightarrow \text{(known for now)}$$

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

Also, let $U_j(t-1, \delta) = \infty$ if $T_j(t-1) = 0$.

(this forces the player to play each arm atleast once initially).

UCB(δ) algorithm:

At each time t: $A_t = \underset{1 \le j \le k}{\operatorname{argmax}} \; U_j(t-1, \delta)$,

Intuition

* Consider the best arm, (arm 1 with $\mu_1 > \mu_2 \ge \mu_3 \cdots$)

for t large, hopefully the player has played this arm a large number of times, i.e.,

$$\left| T_1(t-1) - (t-1) \right| \text{ is small.}$$

$$\therefore \quad U_1(t-1) = \hat{\mu}_1(t-1) + \sqrt{\frac{2\ln(1/\delta)}{T_1(t-1)}}$$

number of samples $T_1(t-1)$ large enough, so $\hat{\mu}_1(t-1) \approx \mu_1$

vanishingly small

$$\therefore \quad U_1(t-1) \approx \mu_1$$

* Consider any other arm $j \geq 2$.

$T_j(t-1)$ is small, but notice that

$\sqrt{\frac{2\ln(1/\delta)}{T_j(t-1)}}$ is a confidence bound

for the estimator $\hat{\mu}_j(t-1)$.

Specifically, if $T_j(t-1)$ were deterministic

estimate with n_j samples $= n_j$ (say)

Then, $(\hat{\mu}_j^{(n_j)} - \mu_j) \sim \frac{1}{\sqrt{n_j}}$ - sub Gaussian.

i.e., $P\left(\hat{\mu}_j^{(n_j)} > \mu_j + \sqrt{\frac{2\ln(1/\delta)}{n_j}}\right) \leq \delta$

This is true for any $n_j \geq 1$!

Choose $\delta = 1/n^2$, and we have that

$U_j(t-1) = \hat{\mu}_j(t-1) + \sqrt{\frac{2\ln(1/\delta)}{T_j(t-1)}}$ is

a high probability upper bound on μ_j.

Thus, these choices of $U_1(t-1)$, $U_j(t-1)$ gives the player the confidence that the TRUE MEAN μ_j is smaller than the TRUE MEAN μ_1 of arm 1 by only comparing $U_1(t-1)$ and $U_j(t-1)$.

In Symbols:

Suppose you observe this

$$\mu_1 \approx U_1(t-1) > U_j(t-1)$$

$$= \hat{\mu}_j(t-1) + \sqrt{\frac{2\ln(1/\delta)}{T_j(t-1)}}$$

$$\text{w.p.} \geq (1-1/n^2) \quad \mu_j$$

$\therefore$ The algorithm choice $A_t = \operatorname{argmax}_j U_j(t-1)$ "ensures" that with high prob:

If $U_1(t-1) > U_j(t-1)$

$\Rightarrow \mu_1 > \mu_j \Rightarrow$ we are indeed playing best arm

The problem in making this rigorous : $T_j(t-1)$

is random, and depends on past samples. However, the anytime concentration result suggests that the above intuition is approximately correct.

Formal Analysis of UCB(δ)

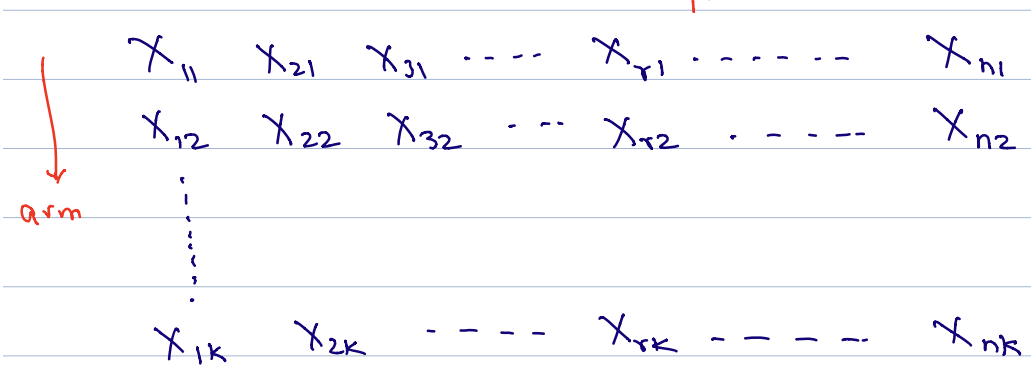
(Thm 7.1 in text).

Thm: With setting as above, and $\delta = 1/n^2$,

$$R_n(\pi, \nu) \leq 3\sum_{i=1}^{k} \Delta_i + \sum_{\{i:\Delta_i>0\}} \left(\frac{16 \ln(n)}{\Delta_i}\right)$$

Proof:

Construct the sequences of r.vs:



where $X_{rj} \sim P_{aj} = \mu_j + 1\text{-subGaussian}$.

There are indep. r.v.s. that will be sampled as needed. For instance, the third time we play arm 5, the r.v. X_{35} is used. In general,

$$X_t = X_{T_{A_t}(t)\, A_t}$$

- X_t: reward at time t
- $T_{A_t}(t)$: The cumulative number of times that arm A_t has been played at time t.
- A_t: the arm index played at time t

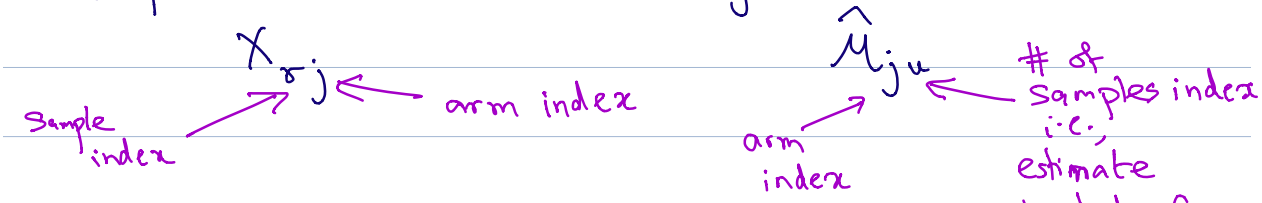
$$R_n = \sum_{j=1}^{k} \Delta_j \, E[T_j(n)] \qquad \text{(Regret Decomp. Lemma)}$$

Fix any arm $j \neq 1$. Define

$$G_j = \left\{ \mu_1 < \min_{1 \le s \le n} U_1(s) \right\} \cap \left\{ \hat{\mu}_{j u_j} + \sqrt{\frac{2}{u_j} \ln\left(\frac{1}{\delta}\right)} < \mu_1 \right\}$$

Note: $\hat{\mu}_{j u_j}$ = empirical estimate for arm j with u_j samples

Side note: I am closely following notation from text, despite minor inconsistencies. E.g.:



Also recall: $\hat{\mu}_j(s)$ — time index (i.e., estimate at time s); arm index

G_j is the good event:

* The UCB index for arm 1 always exceeds the true mean for arm 1

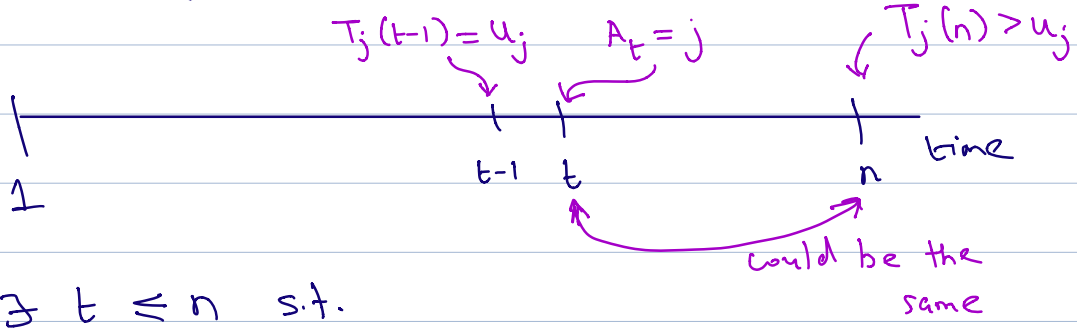
AND

* If the UCB index for arm j was constructed using u_j samples, then this index would be smaller than true mean for arm 1

Two claims for G_j :

(a) If G_j occurs, then $T_j(n) \leq u_j$

Suppose not, i.e., suppose that G_j holds but $T_j(n) > u_j$.



Then $\exists\ t \leq n$ s.t.

$$A_t = j \quad \text{and} \quad T_j(t-1) = u_j .$$

$$U_j(t-1) = \hat{\mu}_j(t-1) + \sqrt{\frac{2\ln(1/\delta)}{T_j(t-1)}}$$

$$= \hat{\mu}_{ju_j} + \sqrt{\frac{2\ln(1/\delta)}{u_j}}$$

$$< \mu_1 \qquad \left(\text{G_j defn}\right)$$

$$< U_1(t-1)$$

Since $U_1(t-1) > U_j(t-1)$, $A_t \neq j$.

This is a contradiction.

(b) $P(G_j^c) \leq n\delta + e^{-u_j \Delta_j^2 / 8} \leq (n+1)\delta$

for $u_j \geq \dfrac{8 \ln(1/\delta)}{\Delta_j^2}$

Recall:

$$G_j^c = \left\{ \mu_1 \geq \min_{1 \leq t \leq n} U_1(t) \right\} \cup \left\{ \hat{\mu}_{j u_j} + \sqrt{\frac{2\ln(1/\delta)}{u_j}} \geq \mu_1 \right\}$$

$$P\left(\mu_1 \geq \min_{1 \leq t \leq n} U_1(t)\right)$$

$$= P\left(\mu_1 \geq \min_{1 \leq t \leq n} \hat{\mu}_1(t-1) + \sqrt{\frac{2\ln(1/\delta)}{T_1(t-1)}}\right)$$

time

t: time
r: samples

note: switching from time to number of samples.

constructed using $T_1(t-1)$ samples

$$\leq \{1, 2, \cdots, n\}$$

$$\leq P\left(\bigcup_{r=1}^{n} \left\{ \mu_1 \geq \hat{\mu}_{1r} + \sqrt{\frac{2\ln(1/\delta)}{r}} \right\} \right)$$

Union bound

# of samples

$$\leq \sum_{r=1}^{n} P\left(\mu_1 \geq \hat{\mu}_{1r} + \sqrt{\frac{2\ln(1/\delta)}{r}}\right)$$

1-subG concentration $\leq \sum_{r=1}^{n} \delta$

$$= n\delta.$$

i.e., $P\left(\mu_1 \geq \min_{1 \leq t \leq n} U_1(t)\right) \leq n\delta$

Next consider the other term in G_j^c:

$$\left\{ \hat{\mu}_{j u_j} + \sqrt{\frac{2\ln(1/\delta)}{u_j}} \geq \mu_1 \right\}$$

Since $u_j \geq \frac{8\ln(1/\delta)}{\Delta_j^2}$, we have

$$\sqrt{\frac{2\ln(1/\delta)}{u_j}} \leq \frac{\Delta_j}{2}$$

$$\therefore P\left(\hat{\mu}_{j,u_j} + \sqrt{\frac{2\ln(1/\delta)}{u_j}} \geq \mu_1\right)$$

$\leq \frac{\Delta_j}{2}$ $\qquad$ $= \mu_j + \Delta_j$

$$\leq P\left(\hat{\mu}_{j,u_j} - \mu_j \geq \frac{\Delta_j}{2}\right)$$

$$\leq e^{-u_j \Delta_j^2 / 8}$$

(1 - SubG concentration)

Thus, for any $u_j \geq \frac{8\ln(1/\delta)}{\Delta_j^2}$, we have

$$P(G_j^c) \leq n\delta + e^{-u_j \Delta_j^2/8} \leq (n+1)\delta$$

Summary: G_j s.t. $T_j(n) \leq u_j$

and $P(G_j^c) \leq n\delta + e^{-u_j\Delta_j^2/8}$

$\leq (n+1)\delta$

for $u_j \geq \frac{8\ln(1/\delta)}{\Delta_j^2}$

$$\therefore \quad E[T_j(n)] = E[\underbrace{T_j(n)}_{\leq u_j} \chi_{G_j}]$$

$$+ E[\underbrace{T_j(n)}_{\leq n} \chi_{G_j^c}]$$

$$\leq u_j + n P(G_j^c)$$

Choose $u_j = \left\lceil \frac{8 \ln(1/\delta)}{\Delta_j^2} \right\rceil$

$$\Rightarrow$$

$$E[T_j(n)] \leq \left\lceil \frac{8 \ln(1/\delta)}{\Delta_j^2} \right\rceil + n(n+1)\delta$$

Choose $\delta = 1/n^2$:

$$E[T_j(n)] \leq \frac{16 \ln(n)}{\Delta_j^2} + 3$$

$$\therefore \quad \Delta_j E[T_j(n)] \leq 3\Delta_j + \frac{16 \ln(n)}{\Delta_j}$$

$$\Rightarrow R_n \leq 3 \sum_{j=1}^{k} \Delta_j + \sum_{j:\Delta_j>0} \frac{16 \ln(n)}{\Delta_j}$$

So far: Instance-dependent bound on regret, i.e., the bound depends on $\{\Delta_j, j=1,2,\cdots,k\}$.

* Bound decays as $\ln(n)$ over the time horizon.

Alternative: Instance-independent bound that does not depend on $\{\Delta_j, j=1,2,\cdots,k\}$ but grows as $\sqrt{n}$ over time.

Thm (7.2 in text): Suppose $\delta = 1/n^2$. Then, for any $\nu \in \mathcal{E}^k_{SG}(1)$,

$$R_n \leq 3 \sum_{j=1}^{k} \Delta_j + 8\sqrt{nk \ln(n)}$$

Proof:

$$R_n = \sum_{j=1}^{k} \Delta_j \, E[T_j(n)]$$

$$\leq 3 + \frac{16 \ln(n)}{\Delta_j^2}$$

$$= \sum_{j=1}^{k} \Delta_j \, E[T_j(n)]$$

from (*) two pages back.

$$\left\{ j : \Delta_j \geq \sqrt{\frac{16k \ln(n)}{n}} \right\}$$

$$+ \sum_{\left\{ j : \Delta_j < \sqrt{\frac{16k \ln(n)}{n}} \right\}} \Delta_j \, E[T_j(n)]$$

$$\leq \sqrt{\frac{16k \ln(n)}{n}}$$

$$\leq \sum_{\left\{ j : \Delta_j \geq \sqrt{\frac{16k \ln(n)}{n}} \right\}} \left(3\Delta_j + \frac{16 \ln(n)}{\Delta_j} \right)$$

$$+ \sqrt{\frac{16k \ln(n)}{n}} \sum_{\left\{ j : \Delta_j < \sqrt{\frac{16k \ln(n)}{n}} \right\}} E[T_j(n)]$$

$$\leq n$$

$$\leq \quad 3 \sum_{j=1}^{k} \Delta_j + \sum_{\{j : \Delta_j \geq \sqrt{\frac{16k\ln(n)}{n}}\}} \left(\frac{16 \ln(n)}{\Delta_j} \right)$$

$$\leq \sqrt{\frac{16k\ln(n)}{n}}$$

$$+ \quad \sqrt{16kn\ln(n)}$$

$$|\cdot| \leq k$$

$$\leq \quad 3 \sum_{j=1}^{k} \Delta_j + k \cdot \sqrt{\frac{n}{k} 16 \ln(n)}$$

$$+ \quad \sqrt{16kn\ln(n)}$$

$$= \quad 3 \sum_{j=1}^{k} \Delta_j + 8\sqrt{nk\ln(n)}$$