UCB bandits w/independent arms

a^* = optimal action

a_t = action selected at time step t

$f^*(a)$ = mean reward of a under true reward

$$\text{regret}(\text{UCB}, \underbrace{T}_{\#\text{ time steps}}) \doteq \sum_{t=1}^{T} \underbrace{f^*(a^*) - f^*(a_t)}_{\substack{\text{difference in}\\ \text{expected best reward}\\ \text{\& expected reward of action selected}}}$$

$$= \sum_{t=1}^{T} \left[U_t(a_t) - f^*(a_t)\right] + \left[f^*(a^*) - U_t(a_t)\right]$$

when $U_t(a_t)$ is an upper confidence bound on the expected reward of action a_t at time t

can use Hoeffding or Azuma inequalities to compute an upper bound on expected reward of each action at each time step

e.g. $U_t(a_t) = \underbrace{\hat{f}_t(a_t)}_{\substack{\text{empirical}\\ \text{avg rew}\\ \text{for action}\\ a_t}} + \sqrt{\dfrac{4\log(t\delta)}{n_t(i)}}$ ← # times tried action a_t — holds w/prob $\geq 1-\delta t^{-2}$

what is the prob that on some step (in T steps) one of the upper confidence bounds will fail to hold?

$$\bigcup_{t=1}^{T} P\left(f^*(a^*) > U_t(a_t)\right) \leq \bigcup_{t=1}^{T} \underbrace{\bigcup_{i=1}^{m}}_{\text{arms}} P\left(\left|f^*(a_i) - \hat{f}(a_i)\right| \geq \sqrt{\frac{4\log(t\delta)}{n_t(i)}}\right)$$

$$\leq \bigcup_{t=1}^{T} \bigcup_{i=1}^{m} \delta t^{-2}$$

aside: $\sum_{i=1}^{\infty} t^{-2} = \frac{\pi^2}{6} < 2$

$$\leq 2m\delta$$

⇒ w/prob at least $1-2m\delta$ UCB will overestimate the true rewards on all T timesteps

⇒ $f^*(a^*) - U_t(a_t) \leq 0$

⇒ $\text{regret}(\text{UCB}, T) \leq \sum_{t=1}^{T} U_t(a_t) - f^*(a_t)$ (dropped 2nd term)

* This is important because now can compare our bounds U_t (which are calculated) to observed reward $f^*(a_t)$ instead of unobserved $f^*(a^*)$

therefore with prob $\geq 1-2m\delta$

$$\text{regret}(UCB, T) \leq \sum_{t=1}^{T} U_t(a_t) - f^*(a_t)$$

$$\leq \sum_{t=1}^{T} \sqrt{\frac{4\log(T\delta)}{n_t(a_t)}}$$

$$\leq \sqrt{4\log(T\delta)} \sum_{t=1}^{T} \sqrt{\frac{1}{n_t(a_t)}}$$

(*) $$= \sqrt{4\log(T\delta)} \sum_{i=1}^{m} \sum_{n=1}^{n_t(i)} \sqrt{1/n}$$ split by each of m actions

note $\sum_{n=1}^{T} \sqrt{1/n} \leq 2\sqrt{T}$ by an integral comparison, then

$\sum_{i=1}^{m} \sum_{n=1}^{n_t(i)} 1/\sqrt{n}$ is maximized if all arms pulled equally

so $$\sum_{i=1}^{m} \sum_{n=1}^{n_t(i)} \frac{1}{\sqrt{n}} \leq \sum_{i=1}^{m} \sum_{n=1}^{T/m} \frac{1}{\sqrt{n}}$$

$$\leq \sum_{i=1}^{m} 2\sqrt{T/m}$$

$$= 2m\sqrt{T/m}$$

$$= 2\sqrt{Tm}$$

substitute this back into (*)

$\text{regret}(UCB, T) \leq 4\sqrt{mT\log(T\delta)}$ with prob at least $1-2m\delta$

can be made tighter in terms of log & constants

- Regret vs Bayesian regret

parameters

$$\text{regret}(\pi, T, \theta) = \sum_{t=1}^{T} E\left[f^*(a^*) - f^*(a_t) \mid \theta\right]$$

$\text{Bayesregret}(\pi, T) = \sum_{t=1}^{T} E\left[f^*(a^*) - f^*(a_t)\right]$ with respect to prior μ over θ

can be related

let $\text{Bayesregret}(\pi, T) = O(g(T))$ for some non-neg func g

then $\forall \epsilon > 0$

$P(\text{Regret}(\pi, T, \theta) \geq g(T)/\epsilon) \leq \epsilon \quad \forall T$ by Markov's inequality

can also use to help bound impact of using wrong prior

- lemma 2 (Posterior sampling). Let ϕ be the distrib of f^* & consider a function g (must satisfy a few properties – see references)

$E[g(f^*) | H_t] = E[g(f_t) | H_t]$ where H_t is the history of observed actions & rewards

$f_t(a_t)$ is our imagined reward

$$\text{regret}(\text{Posterior Sampling}, T) = \sum_{t=1}^{T} f^*(a^*) - f^*(a_t)$$

$$= \sum_{t=1}^{T} |f_t(a_t) - f^*(a_t)| + \underbrace{|f^*(a^*) - f_t(a_t)|}_{=0 \text{ in expec due to lemma}}$$

expec. wrt prior

Posterior sampling bound cont.

$$E\left[\sum_{t=1}^{T} |f_t(x_t) - f^*(x_t)|\right] = E\left[\sum_{t=1}^{T} [f_t(x_t) - U_t(x_t)] + \sum_{t=1}^{T} [U_t(x_t) - f^*(x_t)]\right]$$

$$\leq E\left[\sum_{t=1}^{T} |f_t(x_t) - U_t(x_t)| + \sum_{t=1}^{T} |U_t(x_t) - f^*(x_t)|\right]$$

$$\leq 2E\left[\sum_{t=1}^{T} \underbrace{|U_t(x_t) - L_t(x_t)|}_{\text{so now related to upper \& lower confidence bounds}}\right]$$

$$\text{regret(posterior Sampling, T)} \leq 2E\left[\sum_{t=1}^{T} |U_t(x_t) - L_t(x_t)|\right]$$

$$\leq 2E\left[\sum_{t=1}^{T} 2\sqrt{\frac{4\log(T\delta)}{n_t(a_t)}}\right] \leftarrow \text{using same bounds as UCB}$$

$$\leq 16\sqrt{mT\log(T\delta)} \quad \text{w/prob} \geq 1 - 2m\delta$$

so posterior sampling has same (ignoring constants, logs) regret bounds!