Stochastic Bandits

→ K arms/actions , T rounds

For $t = 1..T$

Algorithm choose a_t action

Receive reward $r_t \in [0,1]$ $\qquad E[r_t(a)] = \mu(a)$

corresponding to action a_t ←

Goal: $\displaystyle\sum_{t=1}^{T} [\mu(a^*) - \mu(a_t)] =: R(T)$ ← explore - exploit tradeoffs

Non-adaptive exploration

(1) Uniform exploration

Try each of the K actions N times ↙ explore

Play highest empirical ~~reward action rest~~ T−NK times ← exploit

$$E[R(T)] = O\left(T^{2/3} (K \log T)^{1/3}\right) \quad \text{for } N \simeq \left(\frac{T}{K}\right)^{2/3} (\log T)^{1/3}$$ ✓

(2) eps-greedy

Toss coin w.p. $\varepsilon_t = t^{-1/3} (K \log t)^{1/3}$ ←

If success chose action uniformly @ random

else choose empirically best action

$$E[R(t)] = O\left(t^{2/3} (K \log t)^{1/3}\right) \quad \text{for any } t \le T.$$

Adaptive Exploration

Action taken ^ depend on previous rounds.

+ no of times its explored

w.h.p $\quad |\hat{\mu}_t(a) - \mu(a)| \le \sigma_t(a) = O\left(\sqrt{\dfrac{\log T}{n_t(a)}}\right)$ ← no. of times action a is taken before t.

$$\mu(a) \in \hat{\mu}_t(a) \pm \sigma_t(a)$$

↑ confidence bound.

$r(a,t)$ $\qquad \begin{bmatrix} r(1,1) & \cdots & r(1,T) \\ \vdots & & \vdots \\ r(K,1) & \cdots & r(K,T) \end{bmatrix}$

Fix n. wp $\geq 1-\delta$ $\qquad |\hat{\mu}(a) - \mu(a)| \leq \sqrt{\frac{\log 1/\delta}{n}}$

$$P(|\quad| \geq \varepsilon) \leq \delta T^2$$

Union bound over all possible actions & all possible values of n.

$\underline{K} \leq \underline{T}$ $\qquad \underline{T}$

$\forall n, \forall a, \quad |\hat{\mu}(a) - \mu(a)| = O\sqrt{\frac{\log 1/\delta}{n_t(a)}} \quad$ wp $\geq 1 - T^2\delta$

$$= O\sqrt{\frac{\log T}{n_t(a)}} =: \sigma_t(a) \qquad \delta = \frac{1}{T^4}$$

(3) Successive elimination

$\rightarrow \hat{\mu}(a) \pm \sigma(a)$

$\mathcal{A} = \{1 .. K\}$

$\rightarrow$ Play all active actions once

Deactivate all a : $\quad UCB_t(a) \leq LCB_t(a')$ for any $a' \in \mathcal{A}$

$\qquad \hat{\mu}_t(a) + \sigma_t(a) \qquad \hat{\mu}_t(a') - \sigma_t(a')$

Note: a^* is always active.

whp. $\quad UCB_t(a^*) = \hat{\mu}_t(a^*) + \sigma_t(a^*)$

$\qquad \geq \mu(a^*)$

$\qquad \geq \mu(a') \geq \hat{\mu}_t(a') - \sigma_t(a') = LCB_t(a')$

If a active $\quad \underbrace{\mu(a^*) - \mu(a)}_{=: \Delta(a)} \leq 2\sigma_t(a^*) + 2\sigma_t(a)$

$\qquad = 4\sigma_t(a)$

$$\because\; n_t(a^*) = \underbrace{n_t(a)}_{\text{active arm @ } t}$$

$$\Delta(a) = O\left(\sqrt{\frac{\log T}{n_T(a)}}\right) \quad \bigstar$$

Let t be last round @ which a is active

$$n_T(a) \leq n_t(a)+1$$

$$R(t) = \sum_{a=1}^{K} n_t(a)\,\Delta(a) = \sum_{a=1}^{K} n_t(a)\, O\left(\sqrt{\frac{\log T}{n_t(a)}}\right)$$

$$\overset{\shortparallel}{\sum_{t=1}^{t} \mu(a^*)-\mu(a_t)}$$

$\sqrt{x}$ is concave

$$= O\left(K\sqrt{\log T}\right) \frac{1}{K}\sum_{a=1}^{K}\sqrt{n_t(a)}$$

$$\underset{\text{Jensen's ineq.}}{\leq} O\left(K\sqrt{\log T}\right)\sqrt{\frac{1}{K}\underbrace{\sum_{a=1}^{K} n_t(a)}_{=\,t}}$$

$$= O\left(\sqrt{Kt\log T}\right) \quad \text{vs.} \quad T^{2/3}(K\log T)^{1/3}$$

Gap-dependent bounds or Instance-based bounds

$$\bigstar \Rightarrow n_T(a) = O\left(\frac{\log T}{(\Delta(a))^2}\right)$$

$$R(T) = \sum_a n_T(a)\,\Delta(a) \leq \sum_a \Delta(a)\, O\left(\frac{\log T}{(\Delta(a))^2}\right)$$

$$= O\left(\sum_a \frac{1}{\Delta(a)} \cdot \log T\right)$$

(4) Upper Confidence Bound (UCB) sampling (Optimism under uncertainty)

pick arm which $\max\limits_a \text{UCB}_t(a)$

$$a_t = \arg\max_a \text{UCB}_t(a)$$

$$\shortparallel$$

$$\hat{\mu}_t(a) + \sigma_t(a)$$

UCB is high either 1) reward is high (exploit)

2) uncertainty is high (explore)

$$\underline{\Delta(a_t)} = \mu(a^*) - \mu(a_t) \le UCB_t(a^*) - (\hat{\mu}(a_t) - \sigma_t(a_t))$$

$$\le UCB_t(a_t) - \hat{\mu}(a_t) + \sigma_t(a_t) \le 2\sigma_t(a_t)$$

from UCB sampling

$$= O\left(\sqrt{\frac{\log T}{n_t(a_t)}}\right) \text{ whp}$$

Same Regret bounds — instance independent.

instance adaptive

Lower Bounds

1. For any bandit algo $\exists$ a problem instance $E[R(T)] = \Omega(\sqrt{KT})$ ←

2. For any bandit algo with non-adaptive exp $\exists$ a problem instance s.t $E[R(T)] = \Omega(K^{1/3} T^{2/3})$

3. For any bandit algo with non-adaptive exp & $E[R(T)] = O(T^r)$ $r \in [2/3, 1)$ for all problem instances then for any problem instance a random permutation of arms yields

$$E[R(T)] = \Omega(T^{\lambda} \sum_a \Delta(a)) \qquad \lambda = 2(1-r)$$

$\times \log T$ $\qquad r \in [\frac{2}{3}, 1)$

4. No algorithm can achieve $E[R(T)] = o(C_I \log t)$ for all problem instances, where C_I - depends on instance I but not on t.

Bandits with prior information

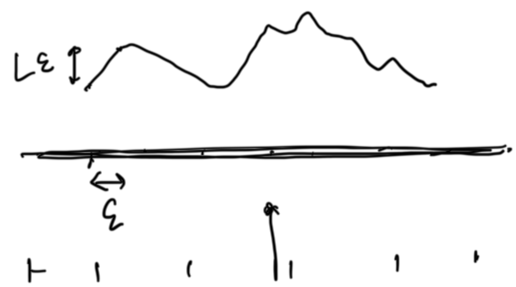
Constrained means — Linear, Lipschitz $\mu \in \mathcal{F}$

holds for continuous arms

Bayesian prior $P(\mu)$

Lipschitz bandits $\quad x \in \mathcal{X} \quad$ continuous arms/actions

$$|\mu(x) - \mu(x')| \le L\,\underbrace{|x - x'|}_{D(x,x')} \qquad \forall\, x, x' \in \mathcal{X} = [0,1]$$



$$E[R(T)] = T\mu(x^*) - \sum_{t=1}^{T} \mu(x_t)$$

$$= \underbrace{T\mu(x^*) - T\mu^*_K}_{\text{discretization error}} + \left(T\mu^*_K - \sum_{t=1}^{T} \mu(x_t)\right)$$

K-armed bandit regret.

best amongst K arms (centers of bins)

$$\le O\left(T\,\frac{L}{K} + \sqrt{KT\log T}\right) \leftarrow \text{adaptive}$$

$$= O\left((L\log T)^{1/3}\, T^{2/3}\right) \qquad K?$$