Formal Setting: MAB

Simplified RL: no states & no transitions

$\mathcal{A}: 1, \dots, K$ "arms"

$r : \mathcal{A} \to \Delta(\mathbb{R})$ noisy $r_t \sim r(a_t)$

$r(r_t | a_t) \in [0, 1]$

$\mathbb{E}(r(a)) = \mu_a$

$T : \mathbb{Z}_+$ integer time horizon

Goal: maximize cumulative expected reward

$$\mathbb{E}\left[\sum_{t=1}^{T} r(a_t)\right] = \sum_{t=1}^{T} \mu_{a_t}$$

Optimal action: $a^* = \underset{a=1,\dots,K}{\operatorname{argmax}} \, \mu_a$

Devise an algorithm for balancing exploration and exploitation

Definition (Regret)

The regret of an algorithm which choosed $a_1, \dots, a_T$

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} r(a^*) - r(a_t)\right]$$

$$= \sum_{t=1}^{T} \mu^* - \mu_{a_t}$$

want to find an algorithm sublinear in regret

$$R(T) \lesssim T^p \quad p < 1$$

$$\lim_{T \to \infty} \frac{R(T)}{T} \to 0 \quad \text{in this case}$$

Balancing exploration & exploitation

Alg 1: Random

for $t=1, \dots, T$:

$$a_t \sim \text{unif}(1, \dots, K)$$

Alg2: Greedy

Try each arm once

compute $\hat{\mu}_{a_t} = r_t$

for $t = K+1, \dots, T$

$$a_t = \arg\max_a \hat{\mu}_a$$

Both suffer from linear regret

Why?

$$R(T) = \sum_{t=1}^{T} \mathbb{E}[r(a^*) - r(a_t)]$$

$$\sum_{t=1}^{T} \mathbb{E}[(r(a^*) - r(a_t)) \mathbb{1}\{a_t \neq a_*\}]$$

$$\sum_{\substack{t:\\ a_t \neq a^*}} \underbrace{(\mu^* - \mu_{a_t})}_{\text{constant wrt } t} \mathbb{P}\{a_t \neq a_*\}$$

$$\geq \sum_{t=1}^{T} \left[\min_{a \neq a^*}(\mu^* - \mu_a)\right] \cdot \mathbb{P}\{a_t \neq a_*\} \geq CT$$

Alg 3: Explore-then-commit

Pull each arm N times ($t = 1, \dots, NK$)
compute $\hat{\mu}_a$ as average observed reward } exploration

for $t = NK+1, \dots, T$

$a_t = \arg\max_a \hat{\mu}_a = \hat{a}^*$ } exploitation (commit)

$$R(T) = \underbrace{\sum_{t=1}^{NK} \mathbb{E}(r(a^*) - r(a_t))}_{R_1} + \underbrace{\sum_{NK+1}^{T} \mathbb{E}(r(a^*) - r(a_t))}_{R_2}$$

Claim: $R_1 \leq NK$ for reward bounded $[0, 1]$