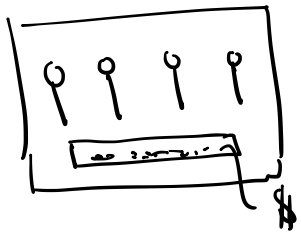
Multi-armed Bandits

MAB

- a gambling machine



← which one to pull?

- don't know which one is the best
- need to <u>explore</u> arms and <u>exploit</u> good ones

- at t, when I play arm j, get <u>reward</u> $X_j(t)$

  ↑ drawn i.i.d. from unknown distn with mean reward μ_j (assume $0 \le X_j(t) \le 1$)

- suffer <u>expected regret</u> $\Delta_j = \mu_* - \mu_j$

  μ_*: mean reward of best arm; μ_j: mean reward for arm I chose

- my current estimate of μ_j:

$$\hat{X}_{j,t} = \underbrace{\frac{1}{T_j(t-1)}}_{\text{\# times I played } j \text{ before } t} \sum_{s=1}^{T_j(t-1)} X_j(s)$$

- total regret for whole game

$$R_n = \sum_{t=1}^{n} \sum_{j=1}^{m} \Delta_j \mathbb{1}_{\{I_t = j\}}$$

t: rounds; Δ_j: regret is Δ_j when we choose arm j at time t

Alg 1: "ε-greedy"

Input: n rounds

m arms

k constant, $k > 10$, $k > \frac{4}{\min_j \Delta_j^2}$ ← difference between best arm & 2nd best arm (not 0)

$\{\varepsilon_t\}_{t=1}^{n} = \min\left\{1, \frac{km}{t}\right\}$ ← 1 when $t < km$, km/t when $t \ge km$

Initialize: play all arms once. $\hat{X}_{j,t} = X_j(j)$ $j=1\dots m$

while $t \le n$

with prob ε_t play an arm uniformly at random "explore" ← choose arm j with prob $1/m$

otherwise (w. prob $1-\varepsilon_t$) play best arm, play j s.t. $\hat{X}_{jt} \ge \hat{X}_{it}$ $\forall i$ "exploit"

get reward $X_j(t)$

update $\hat{X}_{jt}$

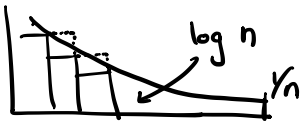
end

Regret bounds are the theory behind MAB algs.

"Theorem 1" $\boxed{\mathbb{E}[R_n] \le O(\log n)}$

Note: $\sum_{t=1}^{n} \frac{1}{t} \le \log n + 1$

$\sum_{t=1}^{n} \frac{1}{t} > \int_1^{n+1} \frac{1}{t}\,dt = \ln(n+1)$



Theorem 1 (Auer, Cesa-Bianchi, Fischer)

$$\mathbb{E}(R_n) \le \sum_{j=1}^{m} \Delta_j \quad \leftarrow \text{initialization}$$

$$+ \sum_{t=m+1}^{n} \sum_{j:\mu_j < \mu_{k^*}} \Delta_j \left(\underbrace{\varepsilon_t \frac{1}{m}}_{O(\frac{1}{t})} + \underbrace{(1-\varepsilon_t)}_{\text{negative}} \underbrace{\beta_j(t)}_{o(1/t)} \right)$$

where $\beta_j(t) = k \left(\frac{t}{mke}\right)^{-k/10} \log\left(\frac{t}{mke}\right) + \frac{4}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-\frac{k\Delta_j^2}{4}}$

(t: can't overpower $t^{-k/10}$; $k/10 > 1$; $\frac{k\Delta_j^2}{4} > 1$)

$\left(\frac{t}{mke}\right)^{-k/10}$: decays faster than $1/t$, $o(1/t)$

$\frac{4}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-\frac{k\Delta_j^2}{4}}$: $o(\frac{1}{t})$

Proof idea:

$$\Delta_j \left(\varepsilon_t \frac{1}{m} + (1-\varepsilon_t) P\left(\hat{X}_{j,T_j(t-1)} \geq \hat{X}_{i,T_i(t-1)} \ \forall i \right) \right)$$

- Δ_j: regret for playing arm j
- ε_t: prob to explore
- $\frac{1}{m}$: prob to play arm j when exploring
- $(1-\varepsilon_t)$: prob to exploit
- $P\left(\hat{X}_{j,T_j(t-1)} \geq \hat{X}_{i,T_i(t-1)} \ \forall i \right)$: prob to play arm j when exploiting. Pr you think arm j is the best when it's not

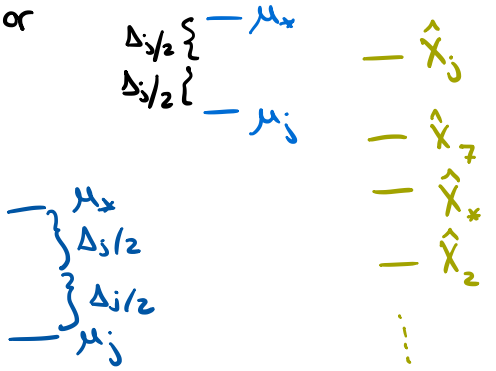
↑ part of proof is to upper bound this by $\mathcal{B}$

Proof: if we think arm j is the best a . c these must have happened:

1) we overestimated μ_j by a lot or
2) we underestimated μ_* by a lot

$$\hat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}$$

$$\hat{X}_{*,T_*(t-1)} \leq \mu_* - \frac{\Delta_j}{2}$$



Step 1:

$$P\left(\hat{X}_{j,T_j(t-1)} \geq \hat{X}_{i,T_i(t-1)} \ \forall i\right) \leq \underbrace{P\left(\hat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right)}_{*1} + \underbrace{P\left(X_{*,T_*(t-1)} \leq \mu_* - \frac{\Delta_j}{2}\right)}_{*2}$$

$P(A \cup B) \leq P(A) + P(B)$

Step 2: Clever Bounding

$$*1 = P\left(\hat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right) = \sum_{s=1}^{t-1} P\left(T_j(t-1) = s, \hat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right)$$

$$= \sum_{s=1}^{t-1} P\left(T_j(t-1) = s \mid \hat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right) \underbrace{P\left(\hat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right)}_{e^{-\Delta_j^2/2 \cdot s} \text{ (Hoeffding)}}$$

$$= \sum_{s=1}^{t-1} P\left(T_j(t-1) = s \,\middle|\, \hat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right) \underbrace{P\left(\hat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right)}_{\text{Hoeffding}}$$

$$\leq e^{-\Delta_j^2/2 \, s}$$

if s is big this term is small

$$X_0 := \frac{1}{2m} \sum_{s'=1}^{t} \varepsilon_{s'}$$

$$= \sum_{s=1}^{\lfloor X_0 \rfloor} \cdots + \sum_{s=\lfloor X_0 \rfloor + 1}^{t-1} \cdots$$

$$\leq e^{-\frac{1}{5} X_0} \qquad \leq \left(\frac{1}{\frac{\Delta_j^2}{2}}\right) e^{-\Delta_j^2/2 \, \lfloor X_0 \rfloor}$$

So far:

$$(*1) \leq e^{-1/5 \, X_0} + \frac{2}{\Delta_j^2} e^{-(\Delta_j/2)\lfloor X_0 \rfloor}$$

$$X_0 \geq \frac{k}{2} \log \frac{t}{mke}$$

$$(*1) \leq \frac{k}{2} \log\left(\frac{t}{mke}\right) \cdot \left(\frac{t}{mke}\right)^{-k/10} + \frac{2}{\Delta_j^2} \left(\frac{t}{mke}\right)^{-k\Delta_j^2/4}$$

Step 3: Turns out, the same bound holds for $(*2)$

Step 4: Combine Steps 1, 2, 3.

$$P\left(\hat{X}_{j,T_j(t-1)} \geq \hat{X}_{i,T_i(t-1)} \;\forall i\right) \leq \quad *2$$

UCB Alg: create an upper confidence bound on each arm
pick the arm with the highest UCB

input: n rounds
m arms

initialize: play all arms once, initialize $\hat{X}_j$ for $j = 1 \dots m$

for $t = m+1 \dots n$

play arm j with highest UCB

$$\hat{X}_{j,T_j(t-1)} + \sqrt{\frac{2\log(t)}{T_j(t-1)}}$$

get reward X_j

update $\hat{X}_j$

end

I I I I

Theorem: regret grows logarithmically in n

$$\mathbb{E}[R_n] \leq \sum_{j=1}^{m} \Delta_j + \sum_{j : \mu_j < \mu_*} \frac{8}{\Delta_j} \log n + \sum_{j=1}^{m} \Delta_j \left(\sum_{t=m+1}^{n} 2t^{-4} (t-1-m)^2 \right)$$