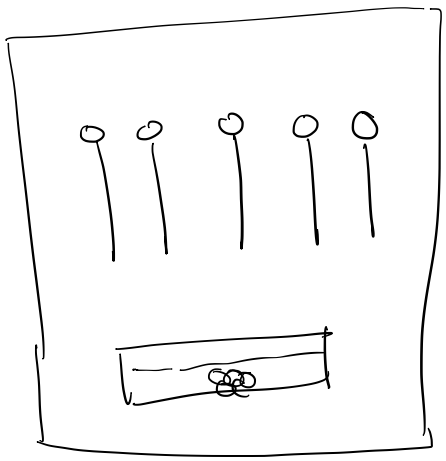
Multi-armed Bandits

MAB

- a gambling machine



- need to explore arms and exploit good ones

- at time t, when I play arm j, get reward $X_j(t)$

  ↑ drawn iid from unknown distn with mean reward μ_j (assume $|X_j(t)| \leq 1$)

- suffer expected regret $\Delta_j = \mu_* - \mu_j$

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

- my current estimate of μ_j:

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

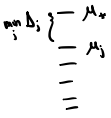
- Total regret for whole game

$$R_n = \sum_{t=1}^{n} \sum_{j=1}^{m} \underbrace{\Delta_j \, \mathbb{1}_{\{I_t = j\}}}_{\text{regret is } \Delta_j \text{ when we chose arm } j \text{ at time } t.}$$

  n: rounds; m: arm

Alg 1 - "ε-greedy"

Input: n rounds
m arms
k constant, $k > 10$ and $k > \frac{4}{\min_j \Delta_j^2}$
$\{\varepsilon_t\}_{t=1}^n = \min\{1, \frac{km}{t}\}$ ← prob to explore

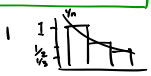


Initialize: play all arms once. $\hat{X}_j = X_j$ $j = 1 \dots m$
while $t < n$
  w.p. ε_t play an arm uniformly at random "explore" ← choose arm j with prob $1/m$
  otherwise (w.p. $1-\varepsilon_t$) play best arm, play j s.t. $\hat{X}_j \geq \hat{X}_i \ \forall i$ "exploit"
  get reward X_j
  update $\hat{X}_j$
end

Regret bounds are the theory behind MAB algs.

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

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



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

Theorem 1 (Auer, Cesa-Bianchi, Fischer)

$$\mathbb{E}(R_n) \leq \sum_{j=1}^{m} \Delta_j + \sum_{t=m+1}^{n} \sum_{j:\mu_j < \mu_*} \Delta_j \left(\varepsilon_t \frac{1}{m} + (1-\varepsilon_t)\beta_j(t) \right)$$

← initialization; $O(t)$; negative

where $\beta_j(t) = k\left(\frac{t}{mke}\right)^{-k/10} \log\left(\frac{t}{mke}\right) + \frac{4}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-\frac{k\Delta_j^2}{4}}$ → 1

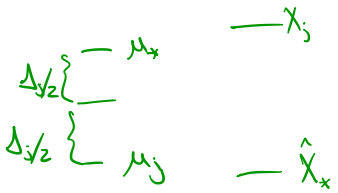
Proof idea:

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

upper bound this

Proof: if we think arm j is the best one of 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: $P\left(\hat{X}_{j,T_j(t-1)} \geq \hat{X}_{i,T_i((t-1)}\right) \leq \underbrace{P\left(\hat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right)}_{*1} + \underbrace{P\left(X_{*,T_*(t-1)} \leq \mu_* - \frac{\Delta_j}{2}\right)}_{*2}$

Step 2: Clever bounding

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

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

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

If s is small, $\uparrow$ could be big...

If s is big we're ok thanks to $e^{-\Delta_j^2/2 \, s}$

s small $\rightsquigarrow$ $s \leq X_0 := \frac{1}{2m} \sum_{s=1}^{\varepsilon} \varepsilon_s$

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

$$= \sum_{s=1}^{\lfloor X_0 \rfloor} P\left(T_j(t-1) = s \,\middle|\, \hat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right) \cdot 1 + \sum_{s=\lfloor X_0 \rfloor}^{t-1} 1 \cdot e^{-(\Delta_j^2/2)\, s}$$

$\leq \frac{1}{\Delta_j^2/2} e^{-\Delta_j^2/2 \lfloor X_0 \rfloor}$ (using $\sum_s e^{-bs} \leq \frac{1}{b} e^{-b}$)

$T_j^{\text{random}}(t-1)$ = # times j is chosen during exploration

$\leq$ ☺

$$\sum_{s=1}^{\lfloor X_0 \rfloor} P\left(T_j^{\text{random}}(t-1) \leq s \,\middle|\, \hat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right)$$

$\perp\!\!\!\perp$ ☺ — T_j^{random} indep of $\hat{X}_{j,s}$

$$\sum_{s=1}^{\lfloor X_0 \rfloor} P\left(T_j^{\text{random}}(t-1) \leq s\right)$$

$\leq$

$$\sum_{s=1}^{\lfloor X_0 \rfloor} P\left(T_j^{\text{random}}(t-1) \leq \lfloor X_0 \rfloor\right)$$

$$\lfloor X_0 \rfloor \, P\left(T_j^{\text{random}}(t-1) \leq \lfloor X_0 \rfloor\right)$$

$\leq$ skipping steps: apply tail bound

$e^{-\gamma_5 \lfloor X_0 \rfloor}$

So far,

$$(*1) \le \lfloor X_0 \rfloor e^{-\frac{1}{5}\lfloor X_0 \rfloor} + \frac{2}{\Delta_j^2} e^{-\left(\frac{\Delta_j^2}{2}\right)\lfloor X_0 \rfloor} \qquad (A)$$

as $\lfloor X_0 \rfloor$ increases ↗ decreases. The worst is if $\lfloor X_0 \rfloor$ is small.

$$X_0 = \frac{1}{2m}\sum_{s=1}^{t} \varepsilon_s = \frac{1}{2m}\sum_{s=1}^{t} \min\left\{1, \frac{km}{s}\right\}$$

$$= \frac{1}{2m}\left[\sum_{s=1}^{\lfloor km \rfloor} 1 + \sum_{s=\lfloor km \rfloor + 1}^{t} \frac{km}{s}\right]$$

$$= \frac{\lfloor km \rfloor}{2m} + \frac{km}{2m}\underbrace{\sum_{s=\lfloor km \rfloor + 1}^{t} \frac{1}{s}}_{VI}$$

$$\log(t+1) - \log(\lfloor km \rfloor + \log e)$$

$$= \frac{k}{2}\frac{\lfloor km \rfloor}{mk} + \frac{k}{2}\log\left(\frac{t+1}{\lfloor km \rfloor e}\right)$$

$$\ge \frac{k}{2}\log\frac{\lfloor km \rfloor}{mk} + \frac{k}{2}\log\left(\frac{t}{\lfloor km \rfloor e}\right)$$

$$= \frac{k}{2}\log\left(\frac{\lfloor km \rfloor}{mk}\frac{t}{\lfloor km \rfloor e}\right) = \frac{k}{2}\log\frac{t}{mke}$$

plug into (A)

$$(*1) \le \frac{k}{2}\log\left(\frac{t}{mke}\right) e^{-\frac{1}{5}\frac{k}{2}\log\frac{t}{mke}} + \frac{2}{\Delta_j^2} e^{-\Delta_j^2/2\left(\frac{k}{2}\log\frac{t}{mke}\right)}$$

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

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

Step 4: Combine Steps 1, 2, 3:

$$P\left(\hat{X}_{j,T_j(t-1)} \ge \hat{X}_{i,T_i(t-1)}\ \forall i\right) \le k\left(\frac{t}{mke}\right)^{-k/10}\log\left(\frac{t}{mke}\right) + \frac{4}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-k\Delta_j^2/4}$$

□

UCB Alg: Create an upper conf 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 $t = m+1 \dots n$

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

Theorem:

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