MULTI-ARM BANDITS

DILEMMA: To EXPLORE or to EXPLOIT

EXPLORE → To explore the settings for parameters

EXPLOIT → To choose the best option.

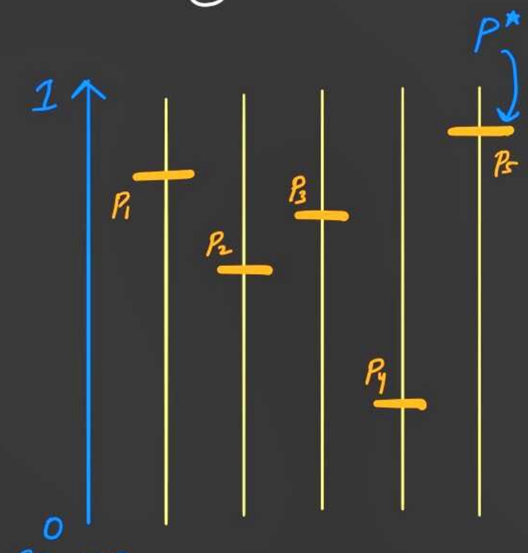
EXAMPLES: ① Online advertising → Template optimisation
② Clinical Trials
③ Packet Routing in a network
④ Game playing and reinforcement learning.

STOCHASTIC MULTI-ARMED BANDITS

$A \rightarrow$ Set of Arms

Arm $a \in A$ has mean reward p_a

$p^* \rightarrow$ Highest mean.



N ARMS with Bernoulli distribution of rewards (Rewards are 0/1)

Pick an arm based on History

DETERMINISTIC ALGORITHM

ALGORITHM: For $t = 0, 1, 2, \ldots T-1$:
- Given history $h^t = \{a^0, r^0, a^1, r^1 \ldots a^{t-1}, r^{t-1}\}$
- pick a^t to sample
- obtain reward r^t

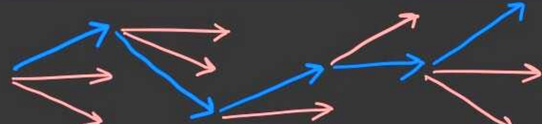
$T \rightarrow$ Horizon / Total Sampling Budget

$$P\{h^T\} = \prod_{t=0}^{T-1} P\{a^t / h^t\} \, P\{r^t / a^t\}$$

$P\{a^t/h^t\}$: Decided by the algorithm

$P\{r^t/a^t\}$: Comes from bandit instance

⊛ NON-DETERMINISTIC ALGORITHM



Parameter $\epsilon \rightarrow$ Controls amount of exploration.

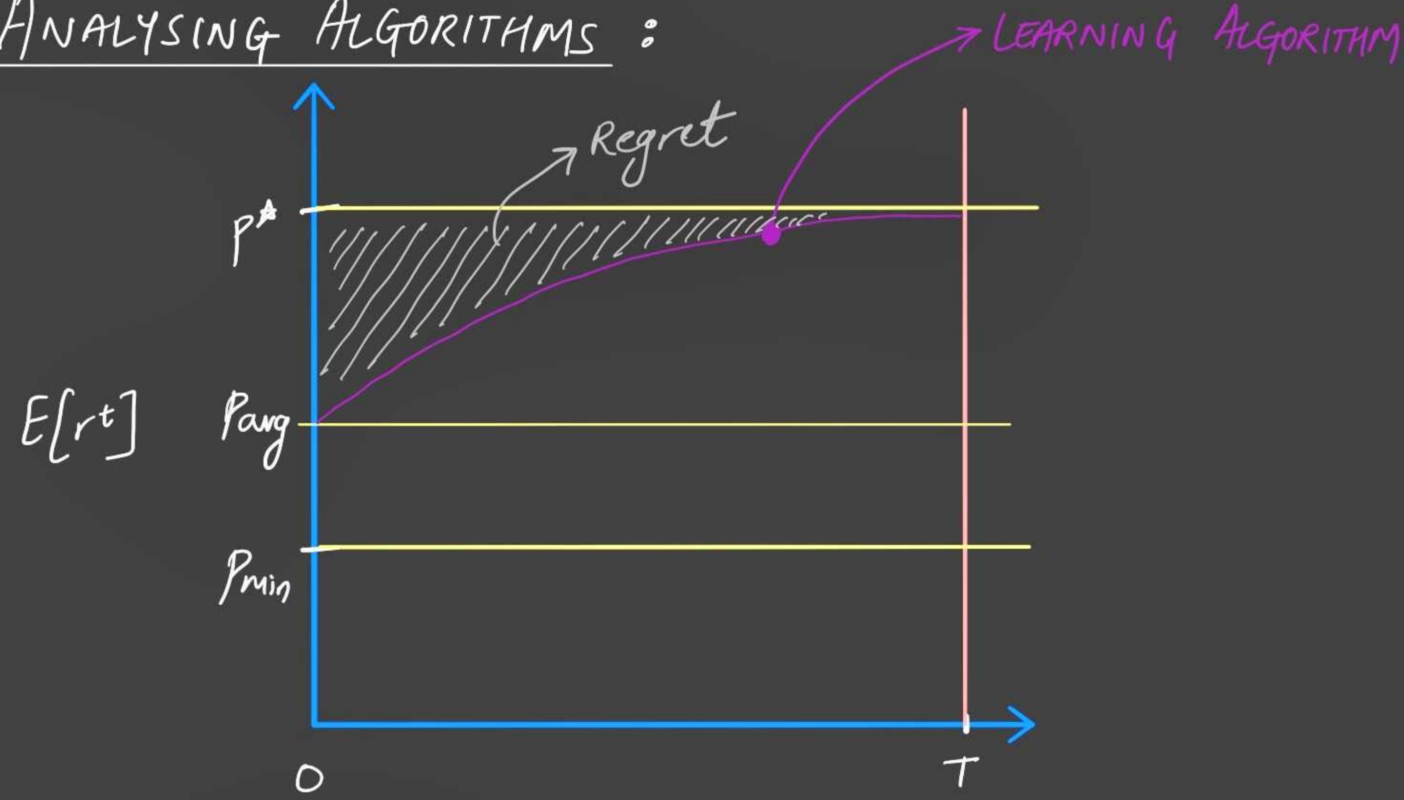
ϵ - GREEDY ALGORITHMS

$\epsilon G1$: - If $t < \epsilon T$, sample an arm uniformly at random
- At $t = \lceil \epsilon T \rceil$, identify a^{best} → Highest empirical mean
- For $t > \epsilon T$, sample a^{best}

$\epsilon G2$:
- if $t < \epsilon T$, sample uniformly
- $t > \epsilon T$, choose arm with highest empirical mean

$\epsilon G3$:
- with probability ϵ, sample an arm uniformly at random, with $1-\epsilon$ sample arm with highest mean

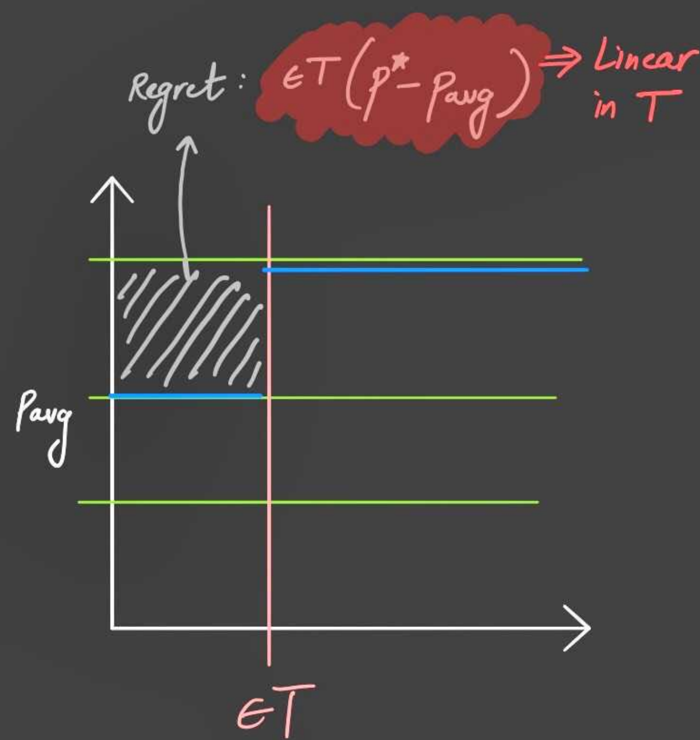
Analysing Algorithms :



Regret $R_T = Tp^* - \sum_{t=0}^{T-1} E[r_t]$

Goal → $\lim_{T\to\infty} \left(\frac{R_T}{T}\right) = 0$ → Sublinear in T

Review of $\epsilon G1$, $\epsilon G2$:



Uniform Sampling

ϵ-first Explore for ϵT pulls and thereafter exploit

Mathematically:

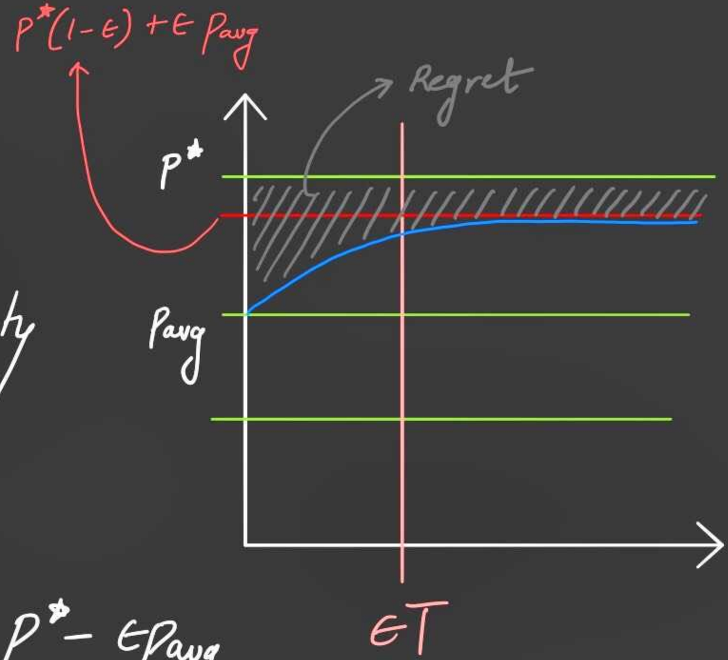
$$R_T = Tp^* - \sum_{t=0}^{T-1} E[r^t] = Tp^* - \sum_{t=0}^{\epsilon T-1} E[r^t] - \sum_{t=\epsilon T}^{T-1} E[r^t]$$

$$= Tp^* - \epsilon T p_{avg} - \sum_{t=\epsilon T}^{T-1} \geq Tp^* - \epsilon T p_{avg} - (T - \epsilon T)p^*$$

$$= \epsilon(p^* - p_{avg})T = \Omega(T) \longrightarrow \text{Linear Regret}$$

Review of $\epsilon G3$:

EXPLORE with probability ϵ, EXPLOIT with probability $1-\epsilon$.



$E[r^t]$ can never exceed $p^* - \epsilon p_{avg}$

Mathematically:

$$R_T = Tp^* - \sum_{t=0}^{T-1} E[r^t] \geq Tp^* - \sum_{t=0}^{T-1} \left((\epsilon) p_{avg} + (1-\epsilon)p^*\right)$$

$$= \epsilon(p^* - p_{avg})T = \Omega(T) \longrightarrow \text{Linear Regret}$$

Conditions for Sub-Linear Regret:

C1. **Infinite Exploration**: As $T \to \infty$, each arm must be pulled infinite number of times

REASON: with probability $(1-p^*)^u$, optimal arm will not be chosen and hence regret will be linear. This is because non-optimal arm will get pulled forever.

ϵ-Greedy satisfy C1

C2. **Greed in the Limit**: Let exploit(T) denote the number of pulls that are greedy with respect to empirical mean upto horizon T.

For sub-linear regret:
$$\lim_{T \to \infty} \frac{E[\text{exploit}(T)]}{T} = 1$$

ϵ-Greedy do not satisfy C2.

Result: An algorithm L achieves sub-linear regret on all instances $I \in \tilde{\mathcal{I}}$ if and only if C1 and C2 are satisfied on all $I \in \tilde{\mathcal{I}}$.

GLIE $\iff$ Sub-linear regret

Greedy Limit

Infinite Exploration

GLIE-fying ϵ-Greedy Strategies:

* ϵ_T-first with $\epsilon_T = \frac{1}{\sqrt{T}}$
* ϵ_t-greedy with $\epsilon_t = \frac{1}{t+1}$

At t^{th} step,

EXPLORE $= \frac{1}{t+1}$

EXPLOIT $= \frac{t}{t+1}$

QUESTION: What if $\epsilon_t = \frac{1}{(t+1)^2}$?

ANSWER: C1 violated, no infinite exploration.

Lower Bound on Regret :

* We desire low regret on all instances

Lai and Robbins

Theorem 2 Let L be an algorithm such that for every bandit instance $I \in \tilde{I}$ and for every $\alpha > 0$, as $T \to \infty$:

$$R_T(L, I) = O(T^\alpha)$$

Then for every bandit instance $I \in \tilde{I}$, as $T \to \infty$:

$$\frac{R_T(L, I)}{\ln(T)} \geq \sum_{a : p(a) \neq p^*(I)} \frac{p^*(I) - p_a(I)}{KL(p_a(I), p^*(I))}$$

KL Divergence on Bernoulli Distribution of these probabilities

* $KL(x, y) \stackrel{def}{=} x \ln\left(\frac{x}{y}\right) + (1-x) \ln\frac{(1-x)}{(1-y)}$ $\quad [0 \ln 0 \stackrel{def}{=} 0]$

Algorithms :

① UCB → Upper Confidence Bounds

- At time t, for every arm a, define $ucb_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln(t)}{u_a^t}}$

Empirical Mean

No of times arm was pulled

- Pull arm having maximal ucb_a^t

REGRET : $O(\log(T))$

② KL-UCB

- At time t, for every arm, define $ucb\text{-}kl_a^t = \max\{q \in [\hat{p}_a^t, 1]$ such that $u_a^t \, KL(\hat{p}_a^t, q) \leq \ln(t) + c\ln(\ln(t))\}$

* $ucb\text{-}kl_a^t \leq ucb_a^t$ → Tighter bound

$\boxed{c \geq 3}$

REGRET: Matches Lai and Robbins lower bound asymptotically

BETA DISTRIBUTION :

$$f(x;\alpha,\beta) = \text{Constant} \cdot x^{\alpha-1}(1-x)^{\beta-1}$$

$$\text{Mean} = \frac{\alpha}{\alpha+\beta} \qquad \text{Variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

③ THOMPSON SAMPLING

- At a time t, let arm a have s_a^t success and f_a^t failures
- $\text{Beta}(s_a^t+1, f_a^t+1) \Rightarrow$ Represents belief about the true mean

$$\text{Mean} = \frac{s_a^t+1}{s_a^t+f_a^t+2} \qquad \text{Variance} = \frac{(s_a^t+1)(f_a^t+1)}{(s_a^t+f_a^t+2)^2(s_a^t+f_a^t+3)}$$

① Computation : For every arm, draw $x_a^t \sim \text{Beta}(s_a^t+1, f_a^t+1)$

② Sampling : Sample arm a for which x_a^t is maximal

REGRET: Matches Lai and Robbins lower bound

CONCENTRATION BOUNDS

① HOEFFDINGS INEQUALITY : $X \in [0,1]$ $E[X]=\mu$

$$P\{\bar{x} \geq \mu+\epsilon\} \leq e^{-2u\epsilon^2} \qquad P\{\bar{x} \leq \mu-\epsilon\} \leq e^{-2u\epsilon^2}$$

② KL INEQUALITY : $x \in [0,1]$ $E[x]=\mu$

Tighter

$$P\{\bar{x} \geq \mu+\epsilon\} \leq e^{-u\,KL(\mu,\mu+\epsilon)} \qquad P\{\bar{x} \leq \mu-\epsilon\} \leq e^{-u\,KL(\mu+\epsilon,\mu)}$$

$$*\,KL(p,q) = p\ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right)$$

ANALYSIS OF UCB ALGORITHM :

TO SHOW: $R_T = O\left(\sum_{a:p_a \neq p^*} \frac{1}{p^* - p_a} \log(T)\right)$

NOTATION :

① $\Delta_a \stackrel{def}{=} p^* - p_a$

② $Z_a^t \longrightarrow$ Event that arm a is pulled at t

③ $z_a^t \longrightarrow$ Random value that has value 1 if arm a is pulled

$\mathbb{E}[z_a^t] = \mathbb{P}\{Z_a^t\}(1) + (1 - \mathbb{P}\{Z_a^t\})(0) = \mathbb{P}\{Z_a^t\}$

④ Instance specific constant

$\bar{u}_a^T \stackrel{def}{=} \left[\frac{8}{(\Delta_a)^2} \ln(T)\right]$

STEP 1 : Show that $R_T = \sum_{a:p_a \neq p^*} \mathbb{E}[u_a^T]\Delta_a$

$$R_T = Tp^* - \sum_{t=0}^{T-1} \mathbb{E}[r^t] = Tp^* - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{P}\{Z_a^t\} \mathbb{E}[r^t | Z_a^t] = Tp^* - \sum_{t=0}^{T-1} \sum_{a \in A} \mathbb{E}[z_a^t] p_a$$

$$= \left(\sum_{a \in A} \mathbb{E}[u_a^T]\right)p^* - \sum_{a \in A} \mathbb{E}[u_a^T] p_a = \sum_{a \in A} \mathbb{E}[u_a^T](p^* - p_a) = \sum_{a \in A} \mathbb{E}[u_a^T]\Delta_a$$

STEP 2 : TWO REGIMES FOR SUBOPTIMAL PULL. TO SHOW: $\mathbb{E}[u_a^T] \leq \bar{u}_a^T + C$ constant

$$\mathbb{E}[u_a^T] = \sum_{t=0}^{T-1} \mathbb{E}[z_a^t] = \underbrace{\sum_{t=0}^{T-1} P\{Z_a^T \wedge (u_a^t < \bar{u}_a^T)\}}_{A} + \underbrace{\sum_{t=0}^{T-1} P\{Z_a^t \wedge (u_a^t \geq \bar{u}_a^T)\}}_{B}$$

STEP 3 : BOUNDING A

$$A = \sum_{t=0}^{T-1} P\{Z_a^t \wedge (u_a^t < \bar{u}_a^T)\} = \sum_{t=0}^{T-1} \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{Z_a^t \wedge (u_a^T = m)\} \leq \sum_{m=0}^{\bar{u}_a^T - 1} 1 = \bar{u}_a^T$$

FLIP SUMMATION

STEP 4 : BOUNDING B

$$B \leq \sum_{t=0}^{T-1} P\left\{\left(\hat{p}_a^T + \sqrt{\frac{2}{u_a^t}\ln(T)} \geq \hat{p}_*^t + \sqrt{\frac{2}{u_*^t}\ln(t)}\right) \wedge (u_a^t \geq \bar{u}_a^T)\right\}$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq \hat{p}_*^t + \sqrt{\frac{2}{y}\ln(t)}\right\}$$

A $\hat{p}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq p_* \Rightarrow \hat{p}_a(x) \geq p_a + \frac{\Delta_a}{2}$

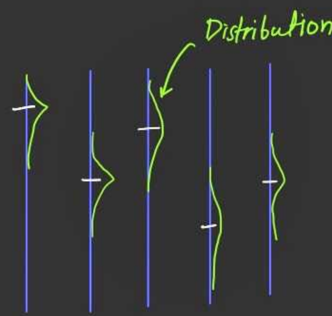
OR

B $\hat{p}_*(y) + \sqrt{\frac{2}{y}\ln(t)} < p_*$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \left(e^{-2x\left(\frac{\Delta_a}{2}\right)^2} + e^{-2y\left(\sqrt{\frac{2}{y}\ln(t)}\right)^2}\right) \leq \sum_{t=0}^{T-1} t^2\left(\frac{2}{t^4}\right) \leq \sum_{t=0}^{\infty} \frac{2}{t^2} = \frac{\pi^2}{3}$$

THOMPSON SAMPLING:

- At a time t, arm has s_a^t successes and f_a^t failures
- $Beta(s_a^t+1, f_a^t+1)$ represents belief about p_a



COMPUTATIONAL STEP: For every arm, draw a sample

$$x_a^t \sim Beta(s_a^t+1, f_a^t+1)$$

SAMPLING STEP: Pull arm which has maximum x_a^t

★ BAYESIAN INFERENCE:

$$P\{A|B\} = P\{B|A\}\, P\{A\} / P\{B\} \qquad Belief_M = P\{w | e_1, e_2 \dots e_M\}$$

$$Belief_{M+1}(w) = P\{w | e_1, e_2 \dots e_{M+1}\} = \frac{P\{e_1, e_2 \dots e_{M+1} | w\}\, P\{w\}}{P\{e_1, e_2 \dots, e_{M+1}\}}$$

CONDITIONAL INDEPENDENCE

$$= \frac{P\{e_1, e_2 \dots e_M | w\}\, P\{e_{M+1}|w\}\, P\{w\}}{P\{e_1, e_2 \dots e_{M+1}\}} = \frac{P\{e_1, e_2 \dots e_M, w\}\, P\{e_{M+1}|w\}}{P\{e_1, e_2 \dots e_{M+1}\}}$$

$$= \frac{P\{w | e_1, e_2 \dots e_M\}\, P\{e_1, e_2, \dots e_M\}\, P\{e_{M+1}|w\}}{P\{e_1, e_2, \dots e_{M+1}\}}$$

$$= \frac{Belief_M(w)\, P\{e_{M+1}|w\}}{\sum_{w' \in W} Belief_M(w')\, P\{e_{M+1}|w'\}}$$

Normalization Constant

CASE 1: $e_{M+1} = 1$ reward

$$Belief_{M+1}(x) = \frac{Belief_M(x) \cdot x}{\int_{y=0}^{1} Belief_M(y)\, y\, dy}$$

CASE 2: $e_{M+1} = 0$ reward

$$Belief_{M+1}(x) = \frac{Belief_M(x) \cdot (1-x)}{\int_{y=0}^{1} Belief_M(y)\,(1-y)\, dy}$$

$$\rightarrow Belief_M(x) = Beta_{s+1, f+1}(x)\, dx.$$

PRINCIPLE OF SELECTING ARM TO PULL:

- We sample a bandit instance I from joint belief distribution and act optimally with respect to I

ALTERNATIVE EXPLANATION: Probability of picking an arm is belief that it is optimal.

OTHER BANDIT PROBLEM :

→ Incorporating risk / variance in the objective
  * RISK MINIMIZATION

→ What if true means vary over time. * Eg. ONLINE ADS
  * → Might take recent data from an interval.

→ Pure Exploration
  * PAC Formulation
  * Simple Regret formulation.

→ Limited number of Feedback
  * $S < T$ Times

→ Large number of arms
  * Quantile Regret ⇒ Look for good and not optimal arms

→ Interaction with many bandits simultaneously.
  * Contextual Bandits

→ Rewards are not from fixed random processes.
  * Adversarial Bandits
  → Necessary to use randomised algorithms