Multi-Armed Bandit

Given unknown probability distributions $R_1, \ldots, R_K$ with means $\mu_1, \ldots, \mu_K$ $\left(\mu^* = \max_{1 \le i \le k} \mu_i\right)$

Choose indices $i_1, i_2, \ldots$ to optimize payout

Regret = difference between cumulative payout and best expectation

$$\rho_T = T \cdot \mu^* - \sum_{t=1}^{T} \hat{r}_t \qquad \hat{r}_t = \text{reward at time } t$$

"optimal" means $P\left(\lim_{T \to \infty} \frac{\rho_T}{T} = 0\right) = 1$ zero regret

Ex:

Arm 1 prob	Arm 1 payout	Arm 2 prob	Arm 2 payout	Arm 3 prob	Arm 3 payout
$\frac{1}{3}$	2	$\frac{1}{4}$	3	$\frac{1}{100}$	200
$\frac{2}{3}$	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{99}{100}$	0
		$\frac{1}{2}$	0		
$\mu_1 = \frac{2}{3}$		$\mu_2 = \frac{7}{8}$		$\mu_3 = 2 = \mu^*$	

uniform rotation: $1\ 2\ 3\ 1\ 2\ 3\ \underbrace{1\ 2\ 3}\ \ldots$

avg regret $2 - \frac{2}{3} = \frac{4}{3}$

avg regret $2 - \frac{7}{8} = \frac{9}{8}$

avg regret $= 0$

tot expected regret over seq of 3 plays $= \frac{4}{3} + \frac{9}{8} = \frac{59}{24}$

$$\lim_{T \to \infty} \frac{\rho_T}{T} = \frac{59}{72}$$

greedy: play each once, play arm w/ best result forever

$1\ 2\ 3\ 1\ 1\ 1\ 1\ 1\ 1 \ldots \quad \lim_{T \to \infty} \frac{\rho_T}{T} = \frac{4}{3}$

$1\ 2\ 3\ 2\ 2\ 2\ 2\ 2\ 2 \ldots \quad \lim_{T \to \infty} \frac{\rho_T}{T} = \frac{9}{8}$

occur with prob > 0

$1\ 2\ 3\ 3\ 3\ 3\ 3\ 3\ 3 \ldots \quad \lim_{T \to \infty} \frac{\rho_T}{T} = 0$

$$P\left(\lim_{T \to \infty} \frac{\rho_T}{T} = 0\right) < 1$$

ε-greedy: play each arm once, then play randomly chosen arm w/ probability ε, arm with best avg observed reward otherwise

$$\lim_{T \to \infty} \frac{\rho_T}{T} = \frac{\varepsilon}{k} \cdot \sum_{i=1}^{k} (\mu^* - \mu_i) > 0$$

(assuming not all arms are optimal)

zero regret

Choose arm j that maximizes $\bar{r}_j + \sqrt{\dfrac{2 \ln T}{n_j}}$

tunable parameter

total plays

avg observed reward for arm j

exploitation

number of times arm j played

exploration

UCB
("upper confidence bound")