K-Armed bandit problem

↓

We face repeatly with a choice among K options and after every choice we receive a Reward from a probability distribution.

GOAL → MAX total reward over some time period

ACTION VALUE: Value of selecting an action a is:

$$q_*(a) = E[R_t \mid A_t = a]$$

How to estimate action value?

1. SAMLE-AVG METHOD

We don't know the reward distribution:

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \, \mathbb{1}_{A_i = a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i = a}} = \frac{\sum R \text{ when take action a}}{\sum \text{ we take action a}}$$

How to select action?

- RANDOM
- GREEDY → Take the max estimated value → SUBOPTIMAL

$$A_t = \arg\max_a Q_t(a)$$

- ε- GREEDY → Behave greedy most of the time but sometimes select randomly from all action wit equal probability

$$A_t = \begin{cases} \arg\max_a Q_t(a) & \text{with prob} = 1-\varepsilon \\ a \sim \mathcal{U}(\{a_1, \dots, a_t\}) & \text{oth.} \end{cases}$$

2.1 INCREMENTAL METHOD (Stationary)

$$Q_n = \frac{R_1 + \dots + R_{n-1}}{n-1} =$$ estimate of the action value of the action that has R_i

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^{n} R_i = \dots = \boxed{Q_n + \frac{1}{n}\left[R_n - Q_n\right]}$$

NewEstimate = OldEstimate + StepSize $\underbrace{[\text{Target} - \text{OldEst}]}_{\text{error estimate}}$

Bandit Algo : $Q(a) = 0 = N(a)$

While True ()

$$A = \begin{cases} \arg\max_a Q(a) & \text{if } P = 1-\varepsilon \\ \text{a random} & \text{else} \end{cases}$$

$R =$ bandit (a) → take an action and return a reward
$N(A) += 1$
$Q(A) = Q(A) + 1/N(A) \left[R - Q(A)\right]$

2.2 INCREMENTAL METHOD (Non Stationary)

Non stationary -> the reward prability changes over time

- Mantain a costant step size:

$$Q_{n+1} = Q_n + \alpha [R_n - Q_n]$$

$\alpha \hookrightarrow (0,1]$

- OPTIMISTIC INITIAL VALUE: Initial action values can be used to IMPROVE EXPLORATION so that the system will do a good amount of exploration.