Recap :- Markov chains.

1. Irreducible
2. Periodicity
3. Recurrent (Positive).

- Ergodic M.C
  1. Irreducible
  2. Aperiodic
  3. Positive recurrent.

- stationary distribution :

π S.D w.r.t M.C with P :-

$\pi \in n \times 1$.

$P \to \underline{n \times n}$.

$$\boxed{\begin{aligned} P\pi &= \pi \\ P^T \pi &= \pi \end{aligned}}$$

$$\left[\pi_1 \ \pi_2 \cdots \pi_n \right] \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & & \cdots & P_{2n} \\ P_{n1} & & & P_{nn} \end{bmatrix} = \underbrace{\left[\pi_1 \ \pi_2 \cdots \pi_n \right]}.$$

P

(*). consider an $\boxed{\text{ergodic M.C}}$ with finite states $\underline{x_1, x_2, \cdots x_n}$.
Define reward function $\underline{R : x \to \mathbb{R}}$.

$R(x_1), R(x_2), \cdots$

$$\lim_{t \to \infty} \frac{E\left[\sum_{n=0}^{t} R(x_n) \right]}{t} \overset{?}{=} \quad \boxed{R(x_n)}$$

x_1

$x_0 = x_1$

$x_1 \sim P(\cdot \mid x_1)$

$\boxed{\mathcal{X} : \text{high} \ \text{medium} \ \text{low}}$

$\rightarrow$ high $\rightarrow$ low $\rightarrow$ medium $\rightarrow$ high $\rightarrow$ ...

$\rightarrow$ 100 50 75 - - -

Realisation of M.C

I 100 + 50 + 75 + - - -

II 50 + 75 + 75 + - - - -

III 75 + 75 + 100 + - - - -

.

.

.

$P(\underline{high} \mid high) = \underline{0.1}$; $P(\underline{low} \mid high) = 0.6$;

$P(\underline{medium} \mid high) = 0.3.$

high, high

high, low

high, medium

high, medium

high, low

.

.

.

$$\lim_{t \to \infty} \frac{E\left[\sum_{n=0}^{t} \underbrace{R(X_n)}\right]}{t}$$

$R(x) \ \forall x$

$$E\left[\sum_{n=0}^{t} R(X_n)\right] = E\left[\sum_{n=0}^{t} \sum_{x \in X} \boxed{\mathbb{1}\{X_n = x\} \cdot R(x)}\right].$$

$x_1, x_2, x_3, \ldots$

$$= E\left[\sum_{x \in X} \sum_{n=0}^{t} \mathbb{1}\{x_n = x\} \cdot R(x)\right]$$

$$= E\left[\sum_{x\in X} \boxed{R(x)} \sum_{n=0}^{t} I\{X_n = x\}\right]$$

$$= \sum_{x\in X} \underline{R(x)} \sum_{n=0}^{t} \underline{Pr\{X_n = x\}}.$$

$$I\{x \in A\} = 1, \text{ if } x \in A$$
$$= 0, \text{ if } x \notin A.$$

$$I\{X_n = x\} = 1, \text{ if } X_n = x$$
$$= 0, \text{ if not}$$

$$\underline{E[I\{X_n = x\}]} = 1 \cdot Pr\{X_n = x\} + 0 \cdot Pr\{X_n \neq x\}$$
$$\rightarrow = \underline{\underline{Pr\{X_n = x\}}}.$$

(*). $$\frac{E\left[\sum_{n=0}^{t} R(X_n)\right]}{t} = \frac{\sum_{x\in X} R(x) \sum_{n=0}^{t} Pr\{X_n = x\}}{t}.$$

$$= \sum_{x\in X} R(x) \cdot \frac{\sum_{n=0}^{t} Pr\{X_n = x\}}{t}$$

(*) $$\boxed{\lim_{t\to\infty} \frac{\sum_{n=0}^{t} Pr\{X_n = x\}}{t} = \underline{\underline{\pi(x)}}. \quad (\#).}$$

$$\therefore \lim_{t\to\infty} \frac{E\left[\sum_{n=0}^{t} R(X_n)\right]}{t} = \underline{\underline{\sum_{\boxed{x\in X}} \underline{R(x)}\, \pi(x)}};$$

□

<u>stationary distn</u> :- $\boxed{\pi P = \pi}$ (1×n, n×n, 1×n). $\underbrace{[\pi(1)\ \pi(2)\ \ldots\ \pi(n)]}$.

1. Basics of Probability
2. M.C

3. optimization
4. Seq. of RVs & convergence

Multi-arm Bandits:-

(*). A bandit with "k" arms.

Pay			might	Reward
50	$\rightarrow$	Arm 1	$\rightarrow$	500 Rs, 0
1	$\rightarrow$	Arm 2	$\rightarrow$	2 Rs, 0
0	$\rightarrow$	Arm 3	$\rightarrow$	1 Rs, 0

$\boxed{X_3} = 1$ or 0

$\rightarrow$ Mode strategy :- Arm 1.

$\rightarrow$ $\dfrac{0+0+0+1+0+\cdots}{n} \rightarrow Pr\{X_3 = 1\}$

$\hookrightarrow E[X_3]$

$\rightarrow$ Mean strategy

" Find the arm with highest expected Value "

$\hookrightarrow$ optimal arm.

(*). Arm $\hookrightarrow$

$\boxed{R}$ is reward you get when arm 1 is pulled.

compute $E[R] \stackrel{?}{=}$

$t=1$; r_1

$t=2$; r_2

$t=3$; r_3

At n-ta $\dfrac{r_n}{n} = E[R]$

$\dfrac{r_n}{n}$:

Naive approach

$r_1, r_2,$

* | r_1 | r_2 | r_3 | $\cdots$ | $r_{100,000}$ |

Running average :-

Q_n : Average of rewards obtained until time 'n-1'.

$$\underline{Q_{n+1}} = \frac{1}{n} \sum_{i=1}^{n} r_i$$

$$= Q_n + \frac{1}{n}\left[r_n - Q_n \right]$$

$\cancel{Q_0 = 0}$

$\cancel{Q_1}$

Q_2

$\vdots$

in general;

$$\boxed{Q_{n+1} = Q_n + \alpha_n \left[R_n - Q_n \right].}$$

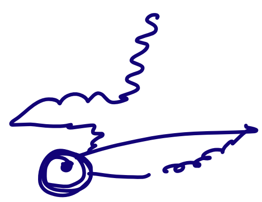
$$\boxed{\sum_{n=1}^{\infty} \alpha_n = \infty} \quad \& \quad \boxed{\sum_{n=1}^{\infty} \alpha_n^2 < \infty.}$$

$\alpha_n > 0$

$$\sum_{n=1}^{\infty} \alpha_n^2$$

Series

α_1^2, $\alpha_1^2 + \alpha_2^2$, $\alpha_1^2 + \alpha_2^2 + \alpha_3^2$, ...



$\{n\}$

1, 2, 3, 4. - - $\to$

① optimal arm : arm with highest expected reward

②: efficient scheme to compute expected values.

Strategies :-

① ϵ-greedy :- 'n' arms.

(t): $Q_t(1)$; $Q_t(2)$ . . $Q_t(n)$.

$Q_t(5)$ is highest.

$t+1$

Pull any random arm w.p. ϵ = 0.0001 (exploration)

Pull best arm so far $1-\epsilon$ (exploitation).

↳ $\arg\max_a Q_t(a)$

"Intelligent exploration"

Arm 1 →

Arm 2 →

UCB :- (action selection scheme) :-

exploration

uncertainty

$$\circledast \quad A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

exploit

↳ no. of times until 't' arm 'a' is pulled.

$Q_t(a)$

$N_t(a)$

I : Arm 1 :- $Q_t(1) \approx Q_t(2)$.

Arm 2 : $N_t(1) << N_t(2)$

Arm 1 why ?

II : Arm 1 : $N_t(1) = N_t(2)$

Arm 2 : $Q_t(1) >> Q_t(2)$

Selection scheme $\propto Q_t(\cdot)$

$\propto \dfrac{1}{N_t(\cdot)}$ ✓