Solutions: Assignment 1

30 July 2025 16:16

MAINAK BISWAS, IISc
PhD student, PMRF.

THEORY

1) :- (8 points)

Markov Chain

(a) $\pi P = \pi$: Stationary distribution π

where π is a row-vector.

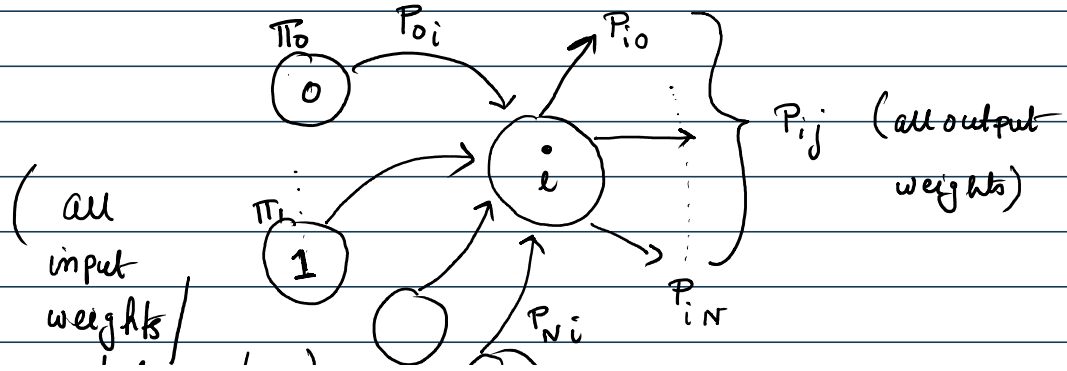
Let $\pi = [\pi_0, \pi_2, \ldots \pi_N] = [\pi_0 \pi_1 \ldots \pi_N] \begin{bmatrix} P_{00} & P_{01} & \ldots & P_{0N} \\ P_{10} & \ldots & \ldots & P_{1N} \\ \ldots & \ldots & \ldots & \ldots \\ & \ldots & \ldots & P_{NN} \end{bmatrix}$

$$\pi_i = \sum_{j=0}^{N} \pi_j P_{ji}$$

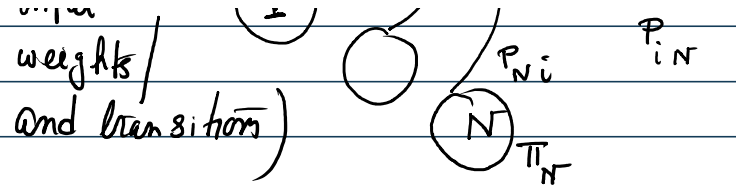
$$\Rightarrow \pi_i (1 - P_{ii}) = \sum_{\substack{j=0 \\ \text{and } j \neq i}}^{N} \pi_j P_{ji}$$

$$1 - P_{ii} = \sum_{j \neq i} P_{ij}$$

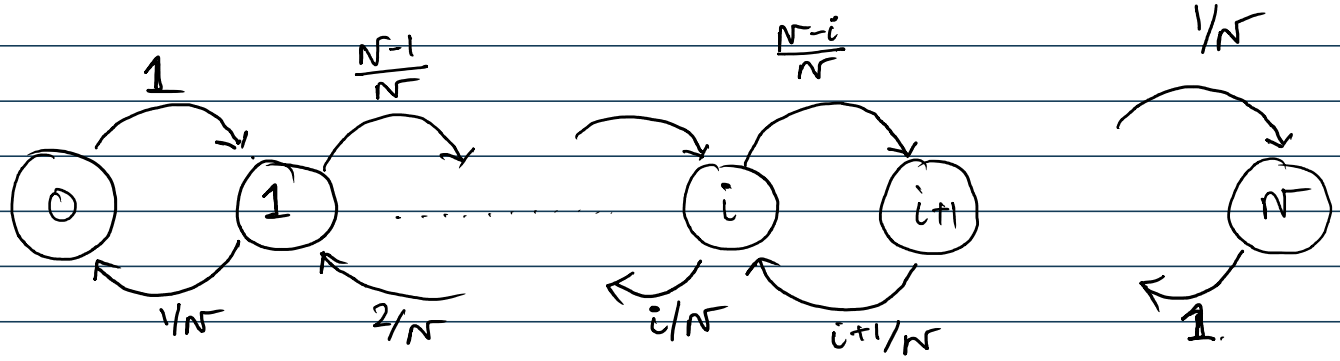
Hence: $$\pi_i \sum_{j \neq i}^{N} P_{ij} = \sum_{\substack{j=0 \\ j \neq i}}^{N} \pi_j P_{ji}$$



weights / and transitions)



(b)



$$P_{xy} = P(X_{t+1} = y \mid X_t = x)$$

$$= \begin{cases} 1 - x/N & y = x+1 \\ x/N & y = x-1 \\ 0 & \text{otherwise.} \end{cases}$$

(c)

Invariant distribution π: (let $\pi = [\pi_0, \pi_1, \ldots, \pi_N]$)

$$\pi P = \pi$$

Using the balance equation from Part (a) $\left[\pi_i \sum_{j \neq i} P_{ij} = \sum_{j \neq i} \pi_j P_{ji}\right]$

State 0: $\pi_0 \cdot \underbrace{(1)}_{\text{(Prob leaving 0)}} = \frac{1}{N}\pi_1 \Rightarrow (\pi_1 = N\pi_0)$

State 1: $\pi_1\left(\frac{N-1}{N} + \frac{1}{N}\right) = \pi_0(1) + \pi_2 \cdot \frac{2}{N}$

$$\Rightarrow \quad \pi_1 = \pi_0 + \frac{2\pi_2}{N}$$

$$\Rightarrow \quad (N-1)\pi_0 = \frac{2\pi_2}{N} \quad \Rightarrow \quad \frac{N(N-1)}{2}\pi_0 = \pi_2$$

State 2:

$$\pi_2 = \pi_1 \cdot \left(\frac{N-1}{N}\right) + \pi_3 \cdot \frac{3}{N}$$

$$\Rightarrow \quad \frac{N(N-1)}{2}\pi_0 = (N-1)\pi_0 + \pi_3 \cdot \frac{3}{N}$$

$$\pi_3 = \frac{N(N-1)(N-2)}{2 \cdot 3}\pi_0 = \frac{N!}{(N-3)!\,3!}\pi_0 = \binom{N}{3}\pi_0$$

$\vdots$

State i: $\quad \pi_i = \dfrac{N(N-1)(N-2)\cdots(N-i+1)\pi_0}{i!} = \dfrac{N!\,\pi_0}{(N-i)!\,i!} = \binom{N}{i}\pi_0$

Law of total prob:

$$\sum_{i=0}^{N} \pi_i = 1$$

$$\Rightarrow \quad \pi_0 \sum_{i=0}^{N} \binom{N}{i} = 1 \qquad \left(\sum_{i=0}^{N}\binom{N}{i} = 2^N\right)$$

$$\pi_0 = \frac{1}{2^N} \quad \Rightarrow \quad \boxed{\pi_i = \frac{1}{2^N}\binom{N}{i}}$$

2) :- (8 pts)

(Multi-armed Bandits)

(a) $\mathbb{P}(R_t \mid A_t = a)$ changes with time.

Given: $\alpha \in [0,1]$

$$Q_{n+1}(a) = Q_n(a) + \underbrace{\alpha}_{\text{(fixed)}} \left[\underbrace{R_{n+1}}_{\text{(reward when action } a \text{ is selected)}} - Q_n(a) \right]$$

$$= (1-\alpha)\, Q_n(a) + \alpha R_{n+1}$$

$$= (1-\alpha)\left\{(1-\alpha)\, Q_{n-1}(a) + \alpha R_n\right\} + \alpha R_{n+1}$$

$$= (1-\alpha)^2 Q_{n-1}(a) + (1-\alpha)\alpha R_n + \alpha R_{n+1}$$

$$= (1-\alpha)^3 Q_{n-2}(a) + (1-\alpha)^2 \alpha R_{n-1} + (1-\alpha)\alpha R_n + \alpha R_{n+1}$$

$$\vdots$$

$$= \alpha \sum_{i=0}^{n} (1-\alpha)^i R_{n-i+1} + (1-\alpha)^{n+1} Q_0(a)$$

Let $(n-i=j)$, $(i = n-j)$

$$= \alpha \sum_{j=0}^{n} (1-\alpha)^{n-j} R_{j+1}$$

$[Q_1 = 0]$ (initialise)

(b) At $(j=0) \rightarrow$ (n steps in the past)

At $(j=0) \rightarrow$ (n steps in the past)

weight : $(1-\alpha)^n$.

at $j=1$;

weight : $(1-\alpha)^{n-1}$ ($n-1$ steps in the past)

exponentially decreasing weightage

$\vdots$

$j=n$ weight : 1 [associated weight with current step]

Thus: the maximum weightage is given to current reward, and it decreases exponentially in the past to discount for rewards from varied reward distribution.

(C) Convergence of Bandits:

$$Q_{n+1} = Q_n + \alpha_n \{ R_{n+1} - Q_n \}$$

$Q_\infty \rightarrow Q_*$ (if step sizes satisfy Robin Munro)

Consider : $\alpha_n = \dfrac{\log(n+2)}{n+2}$ $\quad n \geq 0$.

(i) $\log(n+2) > 0 \quad \forall\, n \geq 0$

$\therefore \alpha_n > 0$. [Condition 1 - positive step sizes]

Condition 2:

(ii) $\sum_n \alpha_n = \infty$ [Asymptotic behaviour of ODE]

To show. $\sum^{\infty} \log(n+2)$ $\rightarrow \infty$

To show: $\sum_{n=0}^{\infty} \frac{\log(n+2)}{n+2} = \infty$

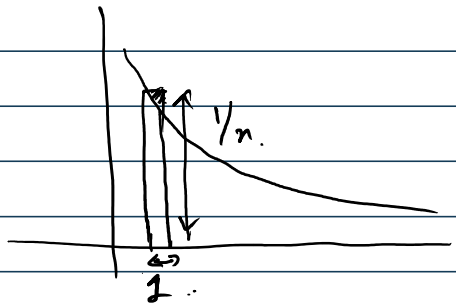
(Consider $\log_2$; else scale accordingly)

$$\frac{1}{n+2} \leq \frac{\log_2(n+2)}{n+2}$$

$\forall\, n \geq 0$; as: $\log_2(n+2) \geq 1$

$$\Rightarrow \sum_{n=0}^{\infty} \frac{1}{n+2} \leq \sum_{n=0}^{\infty} \frac{\log_2(n+2)}{n+2}$$

$$\Rightarrow \sum_{n=0}^{\infty} \frac{1}{n+2} \geq \int_0^{\infty} \frac{1}{n+2}\, dx.$$



$$= \ln(n+2)\Big|_0^{\infty}$$

$$= \infty$$

$$\therefore \quad \infty \leq \sum_{n=0}^{\infty} \frac{\log(n+2)}{n+2}$$

$$\therefore \left(\sum_{n=0}^{\infty} a_n = \infty \right) \quad \text{proved}$$

(iii) $$\sum_{n=0}^{\infty} \left[\frac{\ln(n+2)}{n+2} \right]^2 < \infty$$

Consider the sum: $S_{2K+1} = \sum_{n=2}^{2K+1} \left[\frac{1}{n^p} \right]$

$$\leq 1 + \sum_{n=1}^{k} \left[\frac{1}{(2n)^p} + \frac{1}{(2n+1)^p} \right]$$

$$\leq \frac{1}{(2n)^p}$$

$$\leq 1 + \sum_{n=1}^{k} \frac{2}{(2n)^p}$$

$$= 1 + 2^{1-p} \sum_{n=1}^{k} \frac{1}{n^p}$$

$$S_k \leq S_{2k+1} \quad \text{(more terms)}$$

$$\leq 1 + 2^{1-p} S_{2k+1}$$

$$\Rightarrow S_{2k+1} \leq \frac{1}{1 - 2^{1-p}} < \infty \quad (\forall p > 1) \quad \text{---} *$$

** $\left(\frac{\log x}{x}\right)^2 \leq \frac{1}{x}$.

$\Downarrow$

(won't work.)

(But $\frac{1}{x^{1+\epsilon}}$ $\forall \epsilon > 0$ will)

We know: ** $\log x \leq \sqrt{x}$.

$= \ln x^{(1/2+\delta)}$

Note: $\log x \leq \lim_{\delta \to 0} x^{\frac{1}{2+\delta}}$ $\quad 2\left[\frac{1}{2+\delta} - 1\right] = \left[\frac{-(1+\delta)}{2+\delta}\right]^2$

$$\Rightarrow \left(\frac{\log x}{x}\right)^2 \leq \lim_{\delta \to 0} \left(x^{\frac{-(2+\delta)}{2+\delta} - \frac{\delta}{2+\delta}} \right)$$

$$\leq \frac{1}{x^{1+\epsilon}} \quad (\exists \text{ some } \epsilon)$$

$$\therefore \sum_{x=2}^{\infty} \left(\frac{\log x}{x}\right)^2 \leq \sum_{x=2}^{\infty} \frac{1}{x^{1+\epsilon}} < \infty$$

(from *)

Q.E.D

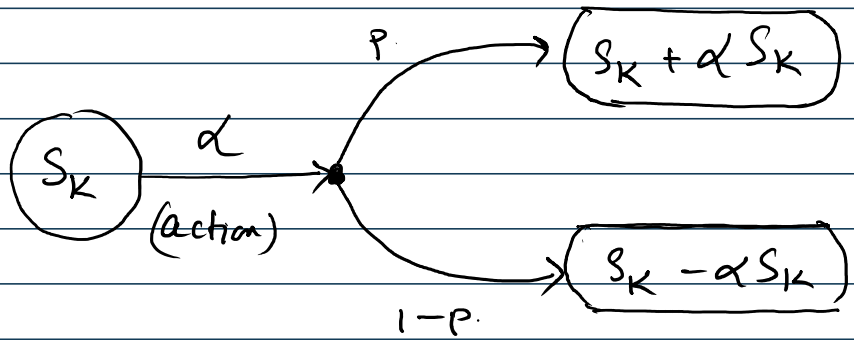
Prove: $\alpha_n = \left(\frac{\log n}{n}\right)^2$ is a valid step schedule.

3) Finite Horizon MDPs; Model Based Method: (8 pts)

Prob. of winning: p
Prob. of losing: $q = 1 - p$
} (at every step)

(a) Let S_k be the amount of money the gambler has at the end of k round.

Let a_k be the action at time k.



Using DP:

Let V_k be the value function at step k.

$$V_N(S_N) = \begin{cases} \log S_N & S_N > 0 \\ 0 & \text{otherwise.} \end{cases}$$

(terminal reward)

$S_{N-1} \rightarrow S_N$ (need $V_0(S_0) = \mathbb{E}\left[\sum_{i=0}^{N} R_i \mid S_0 = s_0\right]$

$$V_{N-1}(S_{N-1}) = \mathbb{E}_{S \sim P(\cdot | S_{N-1}, \alpha_{N-1})}\left[\overrightarrow{R_N}^{0} + V_N(S) \middle| S_{N-1}, \alpha \right]$$

$$= \sup_{\alpha_{N-1}} \left[p\, V_N\left((1+\alpha_{N-1}) S_{N-1} \right) + q\, V_N\left((1-\alpha_{N-1}) S_{N-1} \right) \right]$$

$$= \sup_{\alpha_{N-1}} \left[p \log (1+\alpha_{N-1}) S_{N-1} + (1-p) \log (1-\alpha_{N-1}) S_{N-1} \right]$$

(b)

$$\frac{\partial V_{N-1}}{\partial \alpha_{N-1}} = \left[\frac{p}{(1+\alpha_{N-1})} - \frac{(1-p)}{(1-\alpha_{N-1})} \right] \frac{1}{S_{N-1}} = 0$$

$S_{N-1} > 0$.

$$\Rightarrow \quad \frac{1+\alpha_{N-1}}{1-\alpha_{N-1}} = \frac{p}{1-p}$$

$$\Rightarrow \quad \frac{1+\alpha_{N-1}}{2} = p$$

$$\Rightarrow \quad \alpha_{N-1} = 2p - 1 = 2p - p - q$$

$$\underbrace{(\alpha_{N-1} = p - q)}_{\text{independent of } N}$$

$$\frac{\partial^2 V_{N-1}}{\partial \alpha_{N-1}^2} = -\frac{p}{(1+\alpha_{N-1})^2} - \frac{(1-p)}{(1-\alpha_{N-1})^2} < 0$$

$\alpha_{N-1} = p - q$ (maximizes the profit at step $N-1$)

> 0

$$V_{N-1}(S_{N-1}) = p \log(1+p-q) S_{N-1} + (1-p) \log(1-p+q) S_{N-1}$$

$$= p \log 2p S_{N-1} + (1-p) \log 2q S_{N-1}$$

$$= \log(2q S_{N-1}) + p \log\left(\frac{p}{q}\right)$$

$$= \underbrace{\log 2q + p \log \frac{p}{q}}_{\text{(const independent of } S_{N-1})} + \log S_{N-1}$$

$$= C_{N-1} + \log S_{N-1}$$

Similarly;

$$\therefore V_{N-2}(S_{N-2}) = \sup_{\alpha_{n-2}} \Big[p \{ C_{N-1} + \log((1+\alpha) S_{N-2} \} + (1-p) \{ C_{N-1} \log(1-\alpha) S_{N-2} \} \Big]$$

$$\therefore \frac{\partial V_{N-2}}{\partial \alpha_{N-2}} = 0 \qquad \text{(constants drop off)}$$

$$\Rightarrow \boxed{\alpha_{n-2} = p-q} \quad \text{(independent of } n\text{)}$$

$$\vdots$$

Similarly: strategy for all k : $\boxed{\alpha_k = p-q}$

(c) $p < q$

Now,

$$V_{N-1}(S_{N-1}) = \sup_{\alpha_{n-1}} \left[p \log (1+\alpha_{n-1}) S_{N-1} + (1-p) \log (1-\alpha_{n-1}) S_{N-1} \right]$$

$$= \sup_{\alpha_{n-1}} \left[p \log S_{N-1} + p \log(1+\alpha_{n-1}) + q \log(1-\alpha_{N-1}) + q \log S_{N-1} \right]$$

$$= \sup_{\alpha_{n-1}} \left[\log S_{N-1} + p \log (1+\alpha_{N-1}) + q \log (1-\alpha_{N-1}) \right]$$

$$\frac{\partial V_{N-1}}{\partial \alpha_{n-1}} = \frac{p}{1+\alpha_{N-1}} - \frac{q}{1-\alpha_{N-1}}$$

$$= \frac{(1-\alpha_{N-1})p - (1+\alpha_{N-1})q}{1-\alpha_{N-1}^2}$$

$$= \frac{(p-q) - \alpha_{N-1}(p+q)}{1-\alpha_{N-1}^2}$$

$$= \frac{\overset{<0}{(p-q)} - \overset{\geq 0}{\alpha_{N-1}} \longrightarrow \in [0,1]}{1-\alpha_{N-1}^2} < 0$$

$$\therefore \frac{\partial V_{N-1}}{\partial \alpha_{N-1}} < 0 \quad \text{(decreasing function)}$$

[maximum at $\alpha_{N-1} = 0$] : (Don't bet at any time t)

PROGRAMMING : 16 points)

4) Gradient based Bandits
5) Value Iteration

} Assignment 1. ipynb.