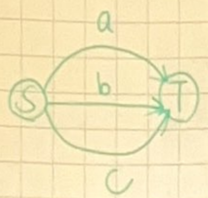
Problem 1:

(a)

① mean vector of $\hat{\nabla}V$ ($E[\hat{\nabla}V]$)



$r_a = 100$ $\quad \theta_a = 0 \quad \pi_\theta(a|s) = \dfrac{e^0}{e^0 + e^{\ln 5} + e^{\ln 4}} = \dfrac{1}{1+5+4} = 0.1$

$r_b = 98$ $\quad \theta_b = \ln 5 \quad \pi_\theta(b|s) = 0.5$

$r_c = 95$ $\quad \theta_c = \ln 4 \quad \pi_\theta(c|s) = 0.4$

$$\log \pi_\theta(a|s) = \log e^{\theta_a} - \log(e^{\theta_a} + e^{\theta_b} + e^{\theta_c})$$

$$\frac{\partial}{\partial \theta_a}[\log \pi_\theta(a|s)] = 1 - \pi_\theta(a|s) = 0.9, \quad \frac{\partial}{\partial \theta_b} = -\pi_\theta(b|s) = -0.5, \quad \frac{\partial}{\partial \theta_c} = -\pi_\theta(c|s) = -0.4$$

$$\log \pi_\theta(b|s) = \log e^{\theta_b} - \log(e^{\theta_a} + e^{\theta_b} + e^{\theta_c})$$

$$\frac{\partial}{\partial \theta_b}[\log \pi_\theta(b|s)] = 1 - \pi_\theta(b|s) = 0.5, \quad \frac{\partial}{\partial \theta_a} = -\pi_\theta(a|s) = -0.1, \quad \frac{\partial}{\partial \theta_c} = -\pi_\theta(c|s) = -0.4$$

$$\log \pi_\theta(c|s) = \log e^{\theta_c} - \log(e^{\theta_a} + e^{\theta_b} + e^{\theta_c})$$

$$\frac{\partial}{\partial \theta_c}[\log \pi_\theta(c|s)] = 1 - \pi_\theta(c|s) = 0.6, \quad \frac{\partial}{\partial \theta_a} = -\pi_\theta(a|s) = -0.1, \quad \frac{\partial}{\partial \theta_b} = -\pi_\theta(b|s) = -0.5$$

$$\Rightarrow \nabla_\theta \log \pi_\theta(a|s) = \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}, \quad \nabla_\theta \log \pi_\theta(b|s) = \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}, \quad \nabla_\theta \log \pi_\theta(c|s) = \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}$$

$$E[\hat{\nabla}V] = \pi_\theta(a|s)[r_a \cdot \nabla_\theta \log \pi_\theta(a|s)] + \pi_\theta(b|s)[r_b \cdot \nabla_\theta \log \pi_\theta(b|s)] + \pi_\theta(c|s)[r_c \cdot \nabla_\theta \log \pi_\theta(c|s)]$$

$$= 0.1 \times 100 \times \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix} + 0.5 \times 98 \times \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix} + 0.4 \times 95 \times \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}$$

$$= \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}$$

NAN PAO

② covariance matrix of $\hat{\nabla} V$

$$E\left[(\hat{\nabla} V - E[\hat{\nabla} V])(\hat{\nabla} V - E[\hat{\nabla} V])^T \right]$$

$$= 0.1 \times \left(\left(100 \times \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}\right) - \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \right) \cdot \left(\left(100 \times \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}\right) - \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \right)^T +$$

$$0.5 \times \left(\left(98 \times \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}\right) - \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \right) \left(\left(98 \times \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}\right) - \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \right)^T +$$

$$0.4 \times \left(\left(95 \times \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}\right) - \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \right) \left(\left(95 \times \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}\right) - \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \right)^T$$

$$= 0.1 \times \begin{bmatrix} 89.7 \\ -50.5 \\ -39.2 \end{bmatrix} \begin{bmatrix} 89.7 & -50.5 & -39.2 \end{bmatrix} + 0.5 \times \begin{bmatrix} -10.1 \\ 48.5 \\ -38.4 \end{bmatrix} \begin{bmatrix} -10.1 & 48.5 & -38.4 \end{bmatrix}$$

$$+ 0.4 \times \begin{bmatrix} -9.8 \\ -48 \\ 57.8 \end{bmatrix} \begin{bmatrix} -9.8 & -48 & 57.8 \end{bmatrix} = \begin{bmatrix} 894.03 & -509.75 & -384.28 \\ -509.75 & 2352.75 & -1843 \\ -384.28 & -1843 & 2227.28 \end{bmatrix}$$ ※

(b) ①

$$B(s) = V^{\pi_\theta}(s) = \sum_a \pi(a|s) Q^{\pi_\theta}(s,a) = 0.1 \times 100 + 0.5 \times 98 + 0.4 \times 95 = 97$$

$$E[\tilde{\nabla} V] = 0.1 \times (100-97) \times \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix} + 0.5 \times (98-97) \times \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix} + 0.4(95-97) \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}$$

$$= \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}$$ ※

② $$E\left[(\tilde{\nabla} V - E[\tilde{\nabla} V])(\tilde{\nabla} V - E[\tilde{\nabla} V])^T \right]$$

$$= 0.1 \times \begin{bmatrix} 2.4 \\ -2 \\ -0.4 \end{bmatrix} \begin{bmatrix} 2.4 & -2 & -0.4 \end{bmatrix} + 0.5 \begin{bmatrix} -0.4 \\ 0 \\ 0.4 \end{bmatrix} \begin{bmatrix} -0.4 & 0 & 0.4 \end{bmatrix}$$

$$+ 0.4 \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix} \begin{bmatrix} -0.1 & 0.5 & -0.4 \end{bmatrix} = \begin{bmatrix} 0.66 & -0.5 & -0.16 \\ -0.5 & 0.5 & 0 \\ -0.16 & 0 & 0.16 \end{bmatrix}$$ ※

(c) find optimal $B(s)$

$$\tilde{\nabla} V - E[\tilde{\nabla} V]$$

for a: $\begin{bmatrix} (100-b)\times 0.9-0.3 \\ (100-b)\times(-0.5)-0.5 \\ (100-b)\times(-0.4)+0.8 \end{bmatrix}$, for b: $\begin{bmatrix} (98-b)\times(-0.1)-0.3 \\ (98-b)\times 0.5-0.5 \\ (98-b)\times(-0.4)+0.8 \end{bmatrix}$, for c: $\begin{bmatrix} (95-b)\times(-0.1)-0.3 \\ (95-b)\times(-0.5)-0.5 \\ (95-b)\times 0.6+0.8 \end{bmatrix}$

$$E\left[(\tilde{\nabla} V - E[\tilde{\nabla} V])(\tilde{\nabla} V - E[\tilde{\nabla} V])^T\right]$$

the trace $= 0.1\times\left[(89.7-0.9b)^2+(-50.5+0.5b)^2+(-39.2+0.4b)^2\right]+$

$0.5\times\left[(-10.1+0.1b)^2+(48.5-0.5b)^2+(-38.4+0.4b)^2\right]+$

$0.4\times\left[(-9.8+0.1b)^2+(-48+0.5b)^2+(57.8-0.6b)^2\right]$

has minimum when $b=97.138$ (set $\frac{d}{db}=0$ to get).

So the optimal $B(s)$ is $\underline{97.138}$ ✱

Problem 2:

Notations: Discount state visitation distribution:

$$d^{\pi}_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s \mid s_0, \pi)$$

$$d^{\pi}_{\mu}(s) = \mathop{E}_{s_0\sim\mu}\left[d^{\pi}_{s_0}(s)\right]$$

$$RHS = \frac{1}{1-\gamma}\mathop{E}_{s\sim d^{\pi_\theta}_{\mu}}\ \mathop{E}_{a\sim\pi_\theta(\cdot|s)}\left[f(s,a)\right]$$

$$= \frac{1}{1-\gamma}\sum_s d^{\pi_\theta}_{\mu}(s)\mathop{E}_{a\sim\pi_\theta(\cdot|s)}\left[f(s,a)\right]$$

$$= \frac{1}{1-\gamma}\sum_s\sum_a \pi_\theta(a|s)\cdot f(s,a)\cdot \underline{d^{\pi_\theta}_{\mu}(s)}$$

$$= \frac{1}{\cancel{1-\gamma}}\mathop{E}_{s_0\sim\mu}\left[\cancel{(1-\gamma)}\sum_s\sum_a \pi_\theta(a|s)f(s,a)\sum_{t=0}^{\infty}\gamma^t P(s_t=s\mid s_0,\pi_\theta)\right]$$

$$= \mathop{E}_{s_0\sim\mu}\left[\sum_s\sum_a \pi_\theta(a|s)\sum_{t=0}^{\infty}\gamma^t P(s_t=s\mid s_0,\pi_\theta)\cdot f(s_t,a_t)\right]$$

$$= \sum_\tau\sum_{t=0}^{\infty} P^{\pi_\theta}_{\mu}(\tau)\,\gamma^t f(s_t,a_t)$$

$$= \mathop{E}_{\tau\sim P^{\pi_\theta}_{\mu}}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t,a_t)\right] = LHS$$

NAN PAO

Problem 3:

Property 1:

consider all possible trajectories:
$$\begin{cases} S\to T & : R_T P_T \\ S\to S\to T & : (R_s+R_T) P_s P_T \\ S\to S\to S\to T & : (R_s+R_s+R_T) P_s^2 P_T \\ \vdots & \vdots \end{cases}$$

$$V(s) = R_T P_T + (R_s+R_T)P_s P_T + (2R_s+R_T)P_s^2 P_T + (3R_s+R_T)P_s^3 P_T + \cdots$$

$$= R_T P_T + R_s P_s P_T + R_T P_s P_T + 2R_s P_s^2 P_T + R_T P_s^2 P_T + 3R_s P_s^3 P_T + R_T P_s^3 P_T + \cdots$$

$$= \frac{R_T P_T}{1-P_s} + R_s P_T \left(P_s + 2P_s^2 + 3P_s^3 + \cdots\cdots \right)$$

$$= \frac{R_T P_T}{1-P_s} + R_s P_T \cdot \sum_{n=1}^{\infty} n P_s^n$$

$$= \frac{R_T P_T}{1-P_s} + \frac{R_s P_T P_s}{(1-P_s)^2} \overset{(1-P_s=P_T)}{=} \frac{R_T P_T}{P_T} + \frac{R_s P_T P_s}{P_T^2} = \frac{P_s}{P_T} R_s + R_T$$

Property 2:

consider all possible trajectories:
$$\begin{cases} S\to T & : R_T P_T \\ S\to S\to T & : \left(\frac{R_s+2R_T}{2}\right) P_T P_s \\ S\to S\to S\to T & : \left(\frac{R_s+2R_s+3R_T}{3}\right) P_T P_s^2 \\ S\to S\to S\to S\to T & : \left(\frac{R_s+2R_s+3R_s+4R_T}{4}\right) P_T P_s^3 \end{cases}$$

R_T

R_s+R_T

$2R_s+R_T$

$3R_s+R_T$

$$V(s) = \sum_{k=0}^{\infty} P_T P_s^k \left(\frac{(1+\cdots+k)R_s + (k+1)R_T}{k+1} \right)$$

$$= \sum_{k=0}^{\infty} P_T P_s^k \left(\frac{\frac{k(k+1)}{2} \cdot R_s + (k+1)R_T}{k+1} \right)$$

$$= \sum_{k=0}^{\infty} P_T P_s^k \cdot \left(\frac{k}{2} R_s + R_T \right)$$

$$= \frac{1}{2} P_T R_s \sum_{k=0}^{\infty} k P_s^k + P_T R_T \sum_{k=0}^{\infty} P_s^k$$

$$= \frac{1}{2} P_T R_s \cdot \frac{P_s}{(1-P_s)^2} + P_T R_T \frac{1}{(1-P_s)}$$

$$= \frac{1}{2} \cancel{P_T} R_s \cdot \frac{P_s}{P_T^{\cancel{2}}} + \cancel{P_T} R_T \cdot \frac{1}{\cancel{P_T}}$$

$$= \frac{P_s}{2P_T} R_s + R_T$$

NAN PAO