HW2

Problem 1

(a)

score function :

$$\frac{d \log\left(\frac{e^{\theta_a}}{e^{\theta_a}+e^{\theta_b}+e^{\theta_c}}\right)}{d\theta_a} = \frac{d\log e^{\theta_a}}{d\theta_a} - \frac{d\log(e^{\theta_a}+e^{\theta_b}+e^{\theta_c})}{d\theta_a}$$

$$= 1 - \pi(a|s)$$

$$\frac{d \log\left(\frac{e^{\theta_a}}{e^{\theta_a}+e^{\theta_b}+e^{\theta_c}}\right)}{d\theta_b} = \frac{d\log e^{\theta_a}}{d\theta_b} - \frac{d\log(e^{\theta_a}+e^{\theta_b}+e^{\theta_c})}{d\theta_b}$$

$$= -\pi(b|s)$$ → action c is similar to b

$$\therefore \nabla_\theta \log \pi_\theta(a|s) = \begin{bmatrix} 1-\pi(a|s) \\ -\pi(b|s) \\ -\pi(c|s) \end{bmatrix}$$

Mean vector :

$$\hat{\nabla} V = \sum_{t=0}^{\infty} \gamma^t \cdot G_t(s_t) \cdot \nabla_\theta \log \pi_\theta(a_t|s_t)$$

$$= r(s_0,a_0) \cdot \nabla_\theta \log \pi_\theta(a_0|s_0)$$

$$E[\hat{\nabla} V] = \sum \pi(\cdot|s) \cdot r(s_0,a_0) \cdot \nabla_\theta \log \pi_\theta(\cdot|s)$$

$$= \frac{1}{e^0+e^{\ln 5}+e^{\ln 4}} \cdot \Big(\overset{X_1}{\underline{e^0 \cdot 100 \cdot \begin{bmatrix} 1-\frac{1}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}}} + \overset{X_2}{\underline{e^{\ln 5} \cdot 98 \cdot \begin{bmatrix} -\frac{1}{10} \\ 1-\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}}}$$

$$+ \overset{X_3}{\underline{e^{\ln 4} \cdot 95 \cdot \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ 1-\frac{4}{10} \end{bmatrix}}} \Big) = \overset{\mu}{\underline{\frac{1}{10} \cdot \left(\begin{bmatrix} 3 \\ 5 \\ -8 \end{bmatrix} \right)}}$$

#

Covariance matrix :

$$E[(\hat{\sigma}V - E[\hat{\sigma}V])\cdot(\hat{\sigma}V - E[\hat{\sigma}V])^T]$$

$$= \frac{1}{10}\left(100\cdot\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)\cdot\left(100\cdot\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)^T +$$

$$\frac{5}{10}\cdot\left(98\cdot\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)\left(98\cdot\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)^T +$$

$$\frac{4}{10}\cdot\left(95\cdot\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)\left(95\cdot\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)^T$$

where $\pi(a\wedge b\mid s_0) = \pi(b\wedge c\mid s_0) = \pi(a\wedge c\mid s_0)\cdot(X_1-\mu)\cdot(X_2-\mu)^T = 0$

$$= \begin{bmatrix}\frac{89403}{100} & \frac{-2039}{4} & \frac{-9607}{25}\\ \frac{-2039}{4} & \frac{9411}{4} & -1843\\ \frac{-9607}{25} & -1843 & \frac{55682}{25}\end{bmatrix} \quad \#$$

(b)

Baseline :

$$V^{\pi_\theta}(s) = \sum_{a_0}\pi(a_0\mid s_0)\cdot Q(s_0,a_0) = \sum_{a_0}\pi(a_0\mid s_0)\cdot r(s_0,a_0)$$

$$= (0.1)\cdot(100) + (0.5)\cdot(98) + (0.4)\cdot(95) = 97$$

Mean vector :

$$E[\hat{\sigma}V] = \frac{1}{10}\cdot\left(1.(100-97)\cdot\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix} + 5.(98-97)\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix}\right.$$

$$\left. + 4.(95-97)\cdot\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix}\right) = \frac{1}{10}\cdot\begin{bmatrix}3\\5\\-8\end{bmatrix} \quad \#$$

Covariance matrix :

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

$$= \frac{1}{10}\left(3\cdot\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)\cdot\left(3\cdot\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)^T +$$

$$\frac{5}{10}\cdot\left(1\cdot\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)\left(1\cdot\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)^T +$$

$$\frac{4}{10}\cdot\left((-2)\cdot\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)\left((-2)\cdot\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\right)^T$$

where $\pi(a \wedge b \mid s_0) = \pi(b \wedge c \mid s_0) = \pi(a \wedge c \mid s_0)\cdot(X_1 - \mu)\cdot(X_2 - \mu)^T = 0$

$$= \begin{bmatrix}0.66 & -0.5 & -0.16\\-0.5 & 0.5 & 0\\-0.16 & 0 & 0.16\end{bmatrix}$$

#

(c)

Suppose baseline $B(s) = b + 97$

Mean vector :

$$E[\hat{\nabla}V] = \frac{1}{10}\cdot\left(1\cdot(3-b)\cdot\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix} + 5\cdot(1-b)\cdot\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix}\right.$$

$$\left. + 4\cdot(-2-b)\cdot\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix}\right) = \frac{1}{10}\cdot\begin{bmatrix}3\\5\\-8\end{bmatrix}$$

Trace of covariance matrix :

$$X_1 - E[\hat{Q}V] = (3-b)\cdot\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix} = \begin{bmatrix}2.4-0.9b\\-2.0+0.5b\\-0.4+0.4b\end{bmatrix}$$

$$X_2 - E[\hat{Q}V] = (1-b)\cdot\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix} = \begin{bmatrix}-0.4+0.1\cdot b\\-0.5b\\0.4+0.4\cdot b\end{bmatrix}$$

$$X_3 - E[\hat{Q}V] = (-2-b)\cdot\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix} - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix} = \begin{bmatrix}-0.1+0.1\cdot b\\0.5+0.5\cdot b\\-0.4-0.6\cdot b\end{bmatrix}$$

$$\begin{aligned}\text{trace} = \frac{1}{1000}\cdot\Big[&1\cdot(24-9b)^2 + 1\cdot(-20+5b)^2 + 1\cdot(-4+4b)^2\\ &+5\cdot(-4+b)^2 + 5\cdot(4+4b)^2 + 5\cdot(-5b)^2\\ &+4\cdot(-1+b)^2 + 4\cdot(5+5b)^2 + 4\cdot(-4-6b)^2\Big]\end{aligned}$$

$$= \frac{1}{1000}\left[580\cdot\left(b-\frac{80}{596}\right)^2 + 1320 - \frac{80^2}{596}\right]$$

$\therefore$ The optimal baseline : $97 + \dfrac{80}{580} \doteqdot 97.138$ #

Problem 2

(a)

P4 under softmax policy : $\dfrac{dV^{\pi_\theta}(\mu)}{d\theta_{s,a}} = \dfrac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a)$

$$\left\| \frac{dV^{\pi_\theta}(\mu)}{d\theta} \right\|_2 = \sqrt{\sum_s \sum_a \left(\frac{dV^{\pi_\theta}(\mu)}{d\theta_{s,a}} \right)^2} \quad (l_2 \text{ norm})$$

$$= \sqrt{\sum_s \sum_a \left(\frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a) \right)^2}$$

$$\geq \underline{\sqrt{\sum_s \left(\frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot A^{\pi_\theta}(s, a^*(s)) \right)^2} \cdot \sqrt{S} \cdot \frac{1}{\sqrt{S}}}$$

by Cauchy-Schwarz inequality

$$\geq \sqrt{\left(\sum_s \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot A^{\pi_\theta}(s, a^*(s)) \right)^2} \cdot \frac{1}{\sqrt{S}}$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|$$

#

(b)

$$\left\| \frac{dV^{\pi_\theta}(\mu)}{d\theta} \right\|_2 \geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \sum_{s'} d_\mu^{\pi^*}(s') \cdot \frac{d_\mu^{\pi_\theta}(s')}{d_\mu^{\pi^*}(s')} \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \sum_{s'} d_\mu^{\pi^*}(s') \cdot \frac{1}{\max_{s'' \in S} \left(\frac{d_\mu^{\pi^*}(s'')}{d_\mu^{\pi_\theta}(s'')} \right)} \cdot \min_{s \in S} \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s', a^*(s')) \right|$$

$$\geq \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min \pi_\theta(a^*(s)|s) \cdot \frac{1}{1-\gamma} \cdot \mathop{E}_{s' \sim d_\mu^{\pi^*}} \left[\mathop{E}_{a' \sim \pi^*} \left[A^{\pi_\theta}(s', a') \right] \right]$$

$$\geq \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min \pi_\theta(a^*(s)|s) \cdot \left(V^*(\mu) - V^{\pi_\theta}(\mu) \right)$$

#

Problem 3

Property 1 :

$$V(s) = \sum_{k=0}^{\infty} P_S^{k} \cdot P_T \cdot (k \cdot R_S + R_T)$$

$$= \frac{P_S}{P_T} \cdot R_S + \frac{R_T}{\cancel{1-P_S}} \cdot \cancel{P_T} = \frac{P_S}{P_T} \cdot R_S + R_T \quad \text{(by lemma 1)}$$

#

Property 2 :

$$E_T[\hat{V}_{MC}(S;T)] = \sum_{k=0}^{\infty} P_S^{k} \cdot P_T \cdot \left(\frac{R_S + \cdots + k \cdot R_S + (k+1) \cdot R_T}{k+1} \right)$$

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

$$= \frac{P_S}{2 \cdot P_T} \cdot R_S + R_T \quad \text{(by lemma 1)}$$

#

lemma 1.

$$f = \sum_{k=0}^{\infty} k \cdot P_S^{k} \cdot R_S \cdot P_T = \sum_{k=0}^{\infty} (k+1) \cdot P_S^{k+1} \cdot R_S \cdot P_T$$

$$-) \quad P_S \cdot f = \sum_{k=0}^{\infty} k \cdot P_S^{k+1} \cdot R_S \cdot P_T$$

$$(1-P_S) \cdot f = \sum_{k=0}^{\infty} P_S^{k+1} \cdot R_S \cdot P_T = \frac{P_S}{\cancel{1-P_S}} \cdot R_S \cdot \cancel{P_T}$$

$$\because P_S + P_T = 1 \qquad \therefore f = \frac{P_S}{P_T} \cdot R_S$$