HW1 - Part 2

Problem 1

$r(s,a) = 100 \qquad \theta_a = 0 \qquad \pi_\theta(a|s) = 0.1$

$r(s,b) = 98 \qquad \theta_b = \ln 5 \qquad \pi_\theta(b|s) = 0.5$

$r(s,c) = 95 \qquad \theta_c = \ln 4 \qquad \pi_\theta(c|s) = 0.4$

(a) What are the $E[\hat{\nabla} V]$ and the convarience matrix of $\hat{\nabla} V$

$$E[\hat{\nabla} V] = E\begin{bmatrix} r(s_t,a_t)\dfrac{\partial \log \pi_\theta(a_t|s_t)}{\partial \theta(s,a)} \\ r(s_t,a_t)\dfrac{\partial \log \pi_\theta(a_t|s_t)}{\partial \theta(s,b)} \\ r(s_t,a_t)\dfrac{\partial \log \pi_\theta(a_t|s_t)}{\partial \theta(s,c)} \end{bmatrix}$$

$$= \begin{bmatrix} \pi_\theta(a|s)\cdot r(s,a)\cdot(1-\pi_\theta(s,a)) - \pi_\theta(b|s)\cdot r(s,b)\,\pi_\theta(s,a) - \pi(c|s)\,r(s,c)\,\pi_\theta(s,a) \\ -\pi_\theta(a|s)\cdot r(s,a)\,\pi_\theta(s,b) + \pi_\theta(b|s)\cdot r(s,b)\,(1-\pi_\theta(s,b)) - \pi(c|s)\cdot r(s,c)\cdot\pi_\theta(s,b) \\ -\pi(a|s)\,r(s,a)\cdot\pi(s,c) - \pi_\theta(b|s)\cdot r(s,b)\cdot\pi_\theta(s,c) + \pi(c|s)\,r(s,c)\cdot(1-\pi_\theta(s,c)) \end{bmatrix}$$

$$= \begin{bmatrix} 0.1\cdot 100\cdot 0.9 - 0.5\cdot 98\cdot 0.1 - 0.4\cdot 95\cdot 0.1 \\ -0.1\cdot 100\cdot 0.5 + 0.5\cdot 98\cdot 0.5 - 0.4\cdot 95\cdot 0.5 \\ -0.1\cdot 100\cdot 0.4 - 0.5\cdot 98\cdot 0.4 + 0.4\cdot 95\, 0.6 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}$$

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

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

$$= E\left[\hat{\nabla}V\hat{\nabla}V^T\right] - \nabla_\theta V^{\pi_\theta} \nabla_\theta V^{\pi_\theta T}$$

$$= E\left[\hat{\nabla}V\hat{\nabla}V^T\right] - \begin{bmatrix} 0.09 & 0.15 & -0.24 \\ 0.15 & 0.25 & -0.4 \\ -0.24 & -0.4 & 0.64 \end{bmatrix}$$

$$\hat{\nabla}V = r(S_0, a_0) \begin{bmatrix} \frac{\partial \log \pi_\theta(a_0|S_0)}{\partial \theta(S,a)} \\ \frac{\partial \log \pi_\theta(a_0|S_0)}{\partial \theta(S,b)} \\ \frac{\partial \log \pi_\theta(a_0|S_0)}{\partial \theta(S,c)} \end{bmatrix}$$

$$\hat{\nabla}V\hat{\nabla}V^T = r(S_0, a_0)^2 \begin{bmatrix} \left(\frac{\partial \log \pi_\theta(a_0|S_0)}{\partial \theta(S,a)}\right)^2 & \frac{\partial\cdots}{\partial\theta(S,a)}\frac{\partial\cdots}{\partial\theta(S,b)} & \frac{\partial\cdots}{\partial\theta(S,a)}\frac{\partial\cdots}{\partial\theta(S,c)} \\ \frac{\partial\cdots}{\partial\theta(S,a)}\frac{\partial\cdots}{\partial\theta(S,b)} & \left(\frac{\partial \log \pi_\theta(a_0|S_0)}{\partial(S,b)}\right)^2 & \frac{\partial\cdots}{\partial\theta(S,b)}\frac{\partial\cdots}{\partial\theta(S,c)} \\ \frac{\partial\cdots}{\partial\theta(S,a)}\frac{\partial\cdots}{\partial\theta(S,c)} & \frac{\partial\cdots}{\partial\theta(S,b)}\cdot\frac{\partial\cdots}{\partial(S,c)} & \left(\frac{\partial \log(a_0|S_0)}{\partial(S,c)}\right)^2 \end{bmatrix}$$

$$E\left[\hat{\nabla}V\hat{\nabla}V^T\right] = \sum \pi(a_0|S_0)\cdot r(a_0 S_0)^2 \cdot \begin{bmatrix} \ddots \end{bmatrix}$$

$$= \begin{bmatrix} 894.12 & -509.6 & -384.52 \\ -509.6 & 2353 & -1843.4 \\ -384.52 & -1843.4 & 2227.92 \end{bmatrix}$$

所求 $= E\left[\hat{\nabla}V\hat{\nabla}V^T\right] - \nabla_\theta V^{\pi_\theta} \nabla_\theta V^{\pi_\theta T} = \begin{bmatrix} 894.03 & -509.75 & -384.28 \\ -509.75 & 2352.75 & -1843 \\ -384.28 & -1843 & 2227.28 \end{bmatrix}$ #

HW1 - Part 2

Problem 1

(b) mean vector of $\hat{\nabla} V$:

$$V^{\pi_\theta}(s) = 0.1 \cdot 100 + 0.5 \cdot 98 + 0.4 \cdot 95 = 97$$

$$\nabla_\theta V^{\pi_\theta}(\mu) \approx \sum_{t=0}^{\infty} \gamma^t \left(G_t - V^{\pi_\theta}(s)\right) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

$$\Rightarrow E[\hat{\nabla} V] = E[(r(s_t,a_t) - V^{\pi_\theta}(s))] \nabla_\theta \log \pi_\theta(a_t|s_t)$$

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

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

$$所求 = E[\hat{\nabla} V \hat{\nabla} V^T] - \nabla_\theta V^{\pi_\theta} \nabla_\theta V^{\pi_\theta T}$$

$$= \sum \pi(a_0|s_0) \cdot \left(r(a_0,s_0) - V^{\pi_\theta}(s)\right)^2 \begin{bmatrix} \left(\frac{\partial \log \pi_\theta(a_0|s_0)}{\partial \theta(s,a)}\right)^2 & \frac{\partial \cdots}{\partial (s,a)} \cdot \frac{\partial \cdots}{\partial (s,b)} & \frac{\partial \cdots}{\partial (s,a)} \cdot \frac{\partial \cdots}{\partial (s,c)} \\ \frac{\partial \cdots}{\partial \theta(s,b)} \cdot \frac{\partial \cdots}{\partial \theta(s,a)} & \left(\frac{\partial \log \pi_\theta(a_0|s_0)}{\partial \theta(s,b)}\right)^2 & \frac{\partial \cdots}{\partial (s,b)} \cdot \frac{\partial \cdots}{\partial (s,c)} \\ \frac{\partial \cdots}{\partial \theta(s,a)} \cdot \frac{\partial \cdots}{\partial \theta(s,c)} & \frac{\partial \cdots}{\partial \theta(s,b)} \cdot \frac{\partial \cdots}{\theta(s,c)} & \left(\frac{\partial \log \pi_\theta(a_0|s_0)}{\partial (s,c)}\right)^2 \end{bmatrix}$$

$$- \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \begin{bmatrix} 0.3, & 0.5, & -0.8 \end{bmatrix}$$

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

Problem 2

(a) Show that $E_{\tau \sim P_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right] = \frac{1}{1-\gamma} E_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)}\left[f(s,a)\right]$

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

$$= \frac{1}{1-\gamma} \sum_s \left(\sum_{s_0} \mu(s_0) (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \cdot P(s_t = s | s_0, \pi_\theta) \right) \cdot \sum_a \pi_\theta(a|s) \cdot f(s,a)$$

$$= \sum_{s_0} \mu(s_0) \cdot \sum_s \sum_a \pi_\theta(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s | s_0, \pi_\theta)\, f(s,a)$$

$$= \sum_{s_0} \mu(s_0) \cdot \sum_s \sum_a \sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = a | s_0, \pi_\theta) \cdot f(s,a)$$

$$= \sum_{s_0} \mu(s_0) \sum_\tau P(\tau | s_0) \cdot \sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)$$

$$= \sum_\tau P(\tau) \sum_{t=0}^{\infty} \gamma^t \cdot f(s_t, a_t)$$

$$= E_{\tau \sim P_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right] = \text{LHS} \quad \#$$

(b) Show that $\nabla_\theta V^{\pi_\theta}(\mu) = E_{\tau \sim P_\mu^{\pi_\theta}}\left[\sum_{t=0}^{T(\tau)-1} \gamma^t A^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$

for episodic environments.

pf: The major difference between episodic and continuing environments is the existence of "terminal state".

Suppose S_* be the terminal state of an episodic environments.

Once the agent reaches S_*, it will stay at S_* forever.

Let $T(\tau)$ be the episodic length of a trajectory τ,

Then,

$$\nabla_\theta V^{\pi_\theta}(\mu) = E_{\tau \sim P_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t, a_t) \cdot \nabla_\theta \log \pi_\theta(a_t|s_t)\right] \quad \cdots\cdots (P5)$$

$$= E_{\tau \sim P_\mu^{\pi_\theta}}\left[\sum_{t=0}^{T(\tau)-1} \gamma^t A^{\pi_\theta}(s_t, a_t) \cdot \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

Problem 3 :

寫程式用數值方法逼近求得最佳的 baseline 約為 97.137931

將式子對 baseline 微分找微分為 0 之地方為 baseline 亦為相近之值

附檔為使用數值方法逼近 baseline 之 code