(1 A).

(a). $\pi_\theta(a|s) = \dfrac{\theta^T x_{s,a}}{\sum_b \theta^T x_{s,b}}$.

$\log \pi_\theta(a|s) = \log \theta^T x_{s,a} - \log \sum_b \theta^T x_{s,b}$.

$\dfrac{1}{\pi_\theta(a|s)} \cdot \nabla \pi_\theta(a|s) = \dfrac{1}{\theta^T x_{s,a}} x_{s,a} - \dfrac{1}{\sum_b \theta^T x_{s,b}} \sum_b x_{s,b}$.

$\therefore \quad \nabla \pi_\theta(a|s) = \pi_\theta(a|s) \left[\dfrac{x_{s,a}}{\theta^T x_{s,a}} - \dfrac{\sum_b x_{s,b}}{\sum_b \theta^T x_{s,b}} \right] =$

(b).

$\ln \pi_\theta(a|s) = \ln\left(\dfrac{1}{\sigma(s,\theta)\sqrt{2\pi}}\right) - \left(\dfrac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right) =$

(i). $\nabla_{\theta_\mu} \ln \pi_\theta(a|s) = \dfrac{(a-\mu(s,\theta))}{\sigma(s,\theta)^2} \cdot x_\mu(s)$.

(ii). $\nabla_{\theta_\sigma} \ln \pi_\theta(a|s) = -\left(\dfrac{1}{\sigma(s,\theta)}\right) \cdot \sigma(s,\theta) \cdot x_\sigma(s) + (a-\mu(s,\theta))^2 \left(\dfrac{1}{\sigma(s,\theta)^3}\right) (\sigma(s,\theta) \cdot x_\sigma(s))$

$= -x_\sigma(s) + \dfrac{(a-\mu(s,\theta))^2}{\sigma(s,\theta)^2} \cdot x_\sigma(s)$.

(2 A).

(a). (H, W) ; (H, S) ; (L, Re) ; (L, W) ; (L, S).

(b). $Q(H,W) \leftarrow (0.9)(0.5) + 0.1[0.5 + 0.5]$

$= 0.55$

$Q(H,S) = (0.9)(0.5) + 0.1[1 + 0.5]$

$= 0.6$

$$Q(L, Re) = (0.9)(0.5) + 0.1\,[0 + 0.6]$$

$$= 0.51$$

$$Q(H, S) = (0.9)(0.6) + 0.1\,[1 + 0]$$

↳ as given in question terminal (state, action) value taken to be 0.

$$= 0.64.$$

(c). $$Q(H, W) = (0.9)(0.5) + 0.1\,[0.5 + (0.6 \times 0.5 + 0.4 \times 0.5)]$$

$$= 0.55$$

$$Q(H, S) = (0.9)(0.5) + 0.1\,[1 + (0.3 \times 0.5 + 0.3 \times 0.5 + 0.4 \times 0.5)]$$

$$= 0.6$$

$$Q(L, Re) = (0.9)(0.5) + 0.1\,[0 + [0.6 \times 0.6 + 0.4 \times 0.55]]$$

$$= 0.45 + 0.1\,[0.58]$$

$$= 0.508$$

$$Q(H, S) = (0.9)(0.6) + 0.1\,[1 + 0]$$

$$= 0.64.$$

(d). $$Q(H, W) = (0.9)(0.5) + (0.1)\,[0.5 + 0.5]$$

$$= 0.55$$

$$Q(H, S) = (0.9)(0.5) + 0.1\,[1 + 0.5]$$

$$= 0.6$$

$$Q(L, Re) = (0.9)(0.5) + 0.1\,[0 + \max\{0.6, 0.55\}]$$

$$= 0.45 + 0.06$$

$$= 0.51.$$

$$Q(H, S) = (0.9)(0.6) + 0.1\,[1 + 0]$$

$$= 0.64.$$

(3 A). (a). From episode 1: $V(1) \leftarrow 0 + \gamma + 0 + \gamma^3 + \cdots = \frac{\gamma}{1-\gamma^2}$

episode 2: $1 + \gamma + \gamma^2 + \cdots = \frac{1}{1-\gamma}$

$\therefore V(1) \leftarrow \frac{\frac{\gamma}{1-\gamma^2} + \frac{1}{1-\gamma}}{2} = \frac{\gamma(1+\gamma)+1}{2(1-\gamma^2)}.$

(b). From episode 1:- every-visit will always yield the infinite seq: 1, 2, 1, 2, ...

$\therefore$ estimate will still be $\frac{\gamma}{1-\gamma^2}$

Similarly for episode 2:- same infinite seq. will be yielded

$\therefore$ estimate is $\frac{1}{1-\gamma}$.

$\therefore V(1) \leftarrow \frac{\gamma(1+\gamma)+1}{2(1-\gamma^2)}$.

(4 A). $V(A) = 0.5\,[V(B)] + 0.5\,[V(C)]$.

(a). $V(B) = 0.5\,[V(D)] + 0.5\,[V(E)]$

$V(C) = 0.5\,[V(D)] + 0.5\,[V(E)]$

$\Big\}$ $V(B) = V(C)$.

$V(D) = 0.5\,[V(B)] + 0.5(1+0). = 0.5\,V(B) + 0.5$

$V(E) = 0.5\,[V(C)] + 0.5(0+0). = 0.5\,V(C)$.

Solving it, we get:

$V(A) = V(B) = V(C) = \frac{1}{2}$; $V(D) = \frac{3}{4}$; $V(E) = \frac{1}{4}$.

(b). $V_1(A) = (0.9)(0.5) + 0.1[0 + 0.5]$.

$= 0.5$

$V_1(C) = (0.9)(0.5) + 0.1[0 + 0.5]$

$= 0.5$

$V_1(E) = (0.9)(0.5) + 0.1[0 + 0]$

$= 0.45$

$V_1(B) = V_1(D) = 0.5$; $V_1(F) = V_1(G) = 0.$