Vicky Zhao

Q1.

$$q_\pi(y, b) \doteq E_\pi\left[\sum_{k=t+1}^{\infty} (R_k - r(\pi)) \mid S_t = y,\ A_t = b \right]$$

$$= E_\pi\left[R_{k+1} - r(\pi) + \sum_{k=t+2}^{\infty} (R_k - r(\pi)) \mid S_t = y,\ A_t = b \right]$$

$$= E_\pi\left[R_{k+1} - r(\pi) + E_\pi\left[\sum_{k=t+2}^{\infty} (R_k - r(\pi)) \mid S_{t+1}, \overset{A_{t+1},\ A_t = b}{S_t = y} \right] \mid S_t = y, A_t = b \right]$$

$$= E_\pi\left[R_{k+1} - r(\pi) + E_\pi\left[\sum_{k=t+2}^{\infty} (R_k - r(\pi)) \mid S_{t+1}, A_{t+1} \right] \mid S_t = y, A_t = b \right]$$

(MP)

$$= E_\pi\left[R_{k+1} - r(\pi) + q_\pi(y_{t+1}, b_{t+1}) \mid S_t = y,\ A_t = b \right]$$

(q_π definition)

$$= \sum_{b, y', r} P_\pi(A_t = b,\ S_t = y',\ R_{t+1} = r \mid S_t = y) \left[r - r(\pi) + q_\pi(y', b) \right]$$

(LOTUS)

$$= \sum_{y', r} p(y', r \mid y, b) \left[r - r(\pi) + \sum_{b'} \pi(b' \mid y')\, q_\pi(y', b') \right]$$

Vicky Zhao

Q2

For the transition S, A, R, S', where the action A~b is drawn from policy distribution b, update the action value like $b=\pi$ (behavior policy = target policy)

following way:

$$Q(S,A) \leftarrow Q(S,A) + \alpha \left[R + \gamma \sum_{a'} \pi(a'|S') Q(S',a') - Q(S,A) \right]$$

using greedification such as ε-greedy.

Vicky Zhao

Q3

1. Switch line 5 and line 6.

2. Add "Choose A' from S' using policy derived from Q" before line 8.

3. Change line 8 with
"$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S',A') - Q(S,A)]$"

4. Add "$A \leftarrow A'$" to line 9.

Vicky Zhao

Q4

$$E_{A\sim b}[\rho(A|S)(R+\gamma V(S')-V(S))\mid S=s]$$

$$= \sum_{a,s',r} P_b(A=a, S'=s', R=r \mid S=s)\,\rho(a|s)(r+\gamma V(s')-V(s))$$

$$= \sum_{a,s',r} b(a|s)\,p(s',r\mid s,a)\,\frac{\pi(a|s)}{b(a|s)}(r+\gamma V(s')-V(s))$$

$$= \sum_{a,s',r} p(s',r\mid s,a)\,\pi(a|s)\,(r+\gamma V(s')-V(s))$$

$$= E_{A\sim\pi}[R+\gamma V(S')-V(S)\mid S=s]$$

On the other hand,

$$E_{A\sim b}[V(S)\mid S=s] = V(s)$$

Therefore, $E_{A\sim b}[V(S)\mid S=s] = S_{A\sim\pi}[V(S)\mid S=s]$.

Vicky Zhao

Q5

$$\frac{\partial g(a)}{\partial a} = g(a)(1-g(a))$$

$$= \frac{1}{1+e^{-a}}\left(1-\frac{1}{1+e^{-a}}\right)$$

$$= \frac{1}{1+e^{-a}} \times \frac{e^{-a}}{1+e^{-a}} = \frac{e^{-a}}{(1+e^{-a})^2}$$

$$g(0) = \frac{1}{1+e^{0}} = \frac{1}{2}$$

$$\frac{\partial \hat{y}_k}{\partial B_{k,j}} = X_j = g(\psi_j) = \frac{1}{1+e^{-\psi_j}}$$

$$\frac{\partial \hat{y}_k}{\partial A_{i,j}} = B_{k,i}\frac{\partial X_i}{\partial A_{i,j}} = B_{k,i}\frac{\partial g(\psi_i)}{\partial \psi_i} S_j$$

$$= B_{k,i}\left(g(\psi_i)(1-g(\psi_i))\right) S_j$$

$$= B_{k,i}\left(\frac{1}{1+e^{-\psi_j}} \times \frac{e^{-\psi_j}}{1+e^{-\psi_j}}\right) S_j$$

$$= B_{k,i} \cdot \frac{e^{-\psi_j}}{(1+e^{-\psi_j})^2} \cdot S_j$$

$$\psi_i = \sum_l A_{i,l} S_l = 0$$

$$\frac{\partial \hat{y}}{\partial B_{k,j}} = g(\psi_i) = g(0) = \frac{1}{2}$$

And $B_{k,i} = 0$ Vicky Zhao

$$\frac{\partial \hat{y}_k}{\partial A_{i,j}} = B_{k,i}\, g'(\psi_i) S_j = 0 \times \left(g(\psi_i) \times (1 - g(\psi_i))\right) S_j$$

$$= 0 \times \left(\frac{1}{2} \times \left(1 - \frac{1}{2}\right)\right) S_j$$

$$= 0 \times \frac{1}{4} \times S_j = 0$$

Vicky Zhao

Q6.

a) Sarsa :

$$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S',A') - Q(S,A)]$$

$$B: Q(S,A) \leftarrow 1.5 + 0.2[0 + 0.9 \times 1.5 - 1.5]$$
$$= 1.5 + 0.2[1.35 - 1.5]$$
$$= 1.5 + 0.2 \times (-0.15)$$
$$= 1.47.$$

$$D: Q(S,A) \leftarrow 1.5 + 0.2[4 + 0.9 \times 0 - 1.5]$$
$$= 1.5 + 0.2[4 + 0 - 1.5]$$
$$= 1.5 + 0.2 \times 2.5$$
$$= 1.5 + 0.5$$
$$= 2.$$

Vicky Zhao

b)

B: $Q(S,A) \leftarrow 1.47 + 0.2[0 + 0.9 \times 2 - 1.47]$

$= 1.47 + 0.2[1.8 - 1.47]$

$= 1.536$

D: $Q(S,A) \leftarrow 2 + 0.2[4 + 0.9 \times 0 - 2]$

$= 2 + 0.2[4 - 2]$

$= 2 + 0.2 \times 2$

$= 2 + 0.4$

$= 2.4$