Assignment 1

Author: **Paolo Renzi**
Contributors:**Bruno Francesco Nocera 1863075, Silverio Manganaro 1817504, Simone Tozzi, 1615930, Leonardo Colosi 1799057, Jacopo Tedeschi 1882789, Amine Ahardane 2050689.**

MARR, RL
November 24, 2023

Theory

Problem 1

given the following Q table:

$$Q(s,a) \;=\; \begin{pmatrix} Q(1,1) & Q(1,2) \\ Q(2,1) & Q(2,2) \end{pmatrix} \;=\; \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

and this parameters:

$$\alpha = 0.1$$
$$\gamma = 0.5$$

and this experience:

$$(s,a,r,s') = (1,2,3,2)$$

Update equation for Q-Learning

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

$$Q(S,A) = 2 + 0.1[3 + (0.5 * 4) - 2] = 2 + (0.1 * 3) = 2.3$$

for Sarsa we are given this next action

$$a' = \pi_\epsilon(s') = 2$$

Update equation for Sarsa

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

$$Q(S,A) = 2 + 0.1[3 + (0.5 * 4) - 2] = 2 + (0.1 * 3) = 2.3$$

Problem 2

We want to prove this:

$$G_{t:t+n} - V_{t+n-1}(S_t) = \sum_{k=t}^{t+n-1} \gamma^k - t\delta_k$$

knowing that:

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... + \gamma^n - 1R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n})$$

So by subtracting $V(S_{t+1})$ to either sides i get

$$\begin{gathered} G_{t:t+n} - V_{t+n-1}(S_t) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... \\ +\gamma^n - 1R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n}) - V_{t+n-1}(S_t) \end{gathered}$$

If we assume that V will not change:

$$= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... + \gamma^n - 1R_{t+n} + \gamma^n V(S_{t+n}) - V(S_t)$$

$$= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... + \gamma^n - 1R_{t+n} + \gamma^n V(S_{t+n}) - V(S_t)$$

Recalling that $\delta_t = R_{t+1}(\gamma V(S_t + 1) - V(s_t))$, we write R_{t+1} in terms of δ_t

$$= \delta_t - \gamma V(S_{t+1}) + V(S_t) + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n}) - V(S_t)$$

Iterate that substitution to all the rewards

$$\begin{gathered} = \gamma^0[\delta_t - \gamma V(S_{t+1}) + V(S_t)] + \gamma[\delta_{t+1} - \gamma V(S_{t+2}) + V(S_{t+1})] + ... \\ + \gamma^{n-1}[\delta_{t+n} - \gamma V(S_{t+n-1}) + V(S_t)] + \gamma^n V(S_{t+n}) - V(S_t) \end{gathered}$$

Write everything in terms of a summation

$$= \sum_{k=t}^{t+n-1} [\gamma^{k-t}\delta_k - \gamma^{k-t+1} V(S_{k+1}) + \gamma^{k-t} V(S_k)] + \gamma^n V(S_{t+n}) - V(S_t)$$

Split the summation in it's components

$$= \sum_{k=t}^{t+n-1} [\gamma^{k-t}\delta_k] - \sum_{k=t}^{t+n-2} [\gamma^{k-t+1} V(S_{k+1})] + \sum_{k=t+1}^{t+n-1} [\gamma^{k-t} V(S_k)]$$

Adjust the indices

$$= \sum_{k=t}^{t+n-1} [\gamma^{k-t}\delta_k] - \sum_{k=t+1}^{t+n-1} [\gamma^{k-t} V(S_k)] + \sum_{k=t+1}^{t+n-1} [\gamma^{k-t} V(S_k)]$$

We get what we wanted to demonstrate

$$= \sum_{k=t}^{t+n-1} \gamma^k - t\delta_k$$

Code

Sarsa(λ)

In Sarsa(λ) I had to implement 2 things:

- The ϵ greedy policy : I did so with an if-then-else, a random number and the max over the Q fuction for the greedy part and sampling randomly the action space for the exploration part
- The update step: I did so by implementig the pseudo-code on the slides (pack 7 slide 71) in particular the equations, being careful on when to use the matrix notation

Q-Learning TD(λ) with RBF

In Q-Learning TD(λ) with RBF I had to implement 2 things:

- The RBF: I did so first trying to implement it myself, but with not so great results, then i tried to use the sklearn implementation but i had to usa also a learned scaler to have good performance, probably because otherwise it would have given too much importance to states that didn't need it
- The update step of Q-Learning TD(λ) : I did so by implementig the pseudo-code on the slides (pack 8 slide 29,) in particular the equations, being careful on when to use the matrix notation