Reinforcement Learning
Assignment 2
Summary

Q1) Exercise 3.4.

Please find the pdf file submitted for Question - 1.

Q2) The final value function of the equiprobable random policy. Comments added in python notebook.

```
values.round(1).reshape(5,5)
```

```
array([[ 3.3,  8.8,  4.4,  5.3,  1.5],
       [ 1.5,  3. ,  2.3,  1.9,  0.5],
       [ 0.1,  0.7,  0.7,  0.4, -0.4],
       [-1. , -0.4, -0.4, -0.6, -1.2],
       [-1.9, -1.3, -1.2, -1.4, -2. ]])
```

Q4) Optimal solutions to the grid world. Comments added in python notebook.

```
print(np.around(value_star, decimals=1))
```

```
[[22.  24.4 22.  19.4 17.5]
 [19.8 22.  19.8 17.8 16. ]
 [17.8 19.8 17.8 16.  14.4]
 [16.  17.8 16.  14.4 13. ]
 [14.4 16.  14.4 13.  11.7]]
```

```
for x in action_star:
    for y in x:
        print(y, end= ',')
    print()
```

```
['>'],['^' '<' 'd' '>'],['<'],['^' '<' 'd' '>'],['<'],
['^' '>'],['^'],['^' '<'],['<'],['<'],
['^' '>'],['^'],['^' '<'],['^' '<'],['^' '<'],
['^' '>'],['^'],['^' '<'],['^' '<'],['^' '<'],
['^' '>'],['^'],['^' '<'],['^' '<'],['^' '<'],
```

Q3)

- Exercise 3.15 and 3.16 (Images also added in the 'Images' folder)

Exercise – 3.15

Adding a constant c to all rewards.

$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} r_{t+1+k}\,\gamma^k \,\middle|\, S_t = s\right]$$

If $r' = r + c$.

then.

new values of states $V'_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} (r_{t+1+k} + c)\,\gamma^k \,\middle|\, S_t = s\right]$

$$= E_\pi\left[\sum_{k=0}^{\infty} r_{t+1+k}\,\gamma^k \,\middle|\, S_t = s\right] + E_\pi\left[\sum_{k=0}^{\infty} c\,\gamma^k \,\middle|\, S_t = s\right]$$

$$= V_\pi(s) + \frac{c}{1-\gamma}$$

$$\therefore V'_\pi(s) = V_\pi(s) + \frac{c}{1-\gamma}$$

$V_c = \frac{c}{1-\gamma}$ $\longrightarrow$ constant added to all states.

It does not affect relative value of states under any policies.

Ex – 3.16

If the task were an episodic one, then,

$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

$$= E_\pi\left[\sum_{k=0}^{T(s)} r_{t+1+k}\,\gamma^k \,\middle|\, S_t = s\right]$$

$T(s)$ is the number of steps, starting from current state until the terminal state is reached.

$$V'_\pi(s) = E_\pi\left[\sum_{k=0}^{T(s)} (r_{t+1+k} + c)\,\gamma^k \,\middle|\, S_t = s\right]$$

$$= E_\pi\left[\sum_{k=0}^{T(s)} r_{t+1+k}\,\gamma^k \,\middle|\, S_t = s\right] + E_\pi\left[\sum_{k=0}^{T(s)} c\,\gamma^k \,\middle|\, S_t = s\right]$$

$$V'_\pi(s) = V_\pi(s) + c\left(\frac{1-(\gamma)^{T(s)}}{1-\gamma}\right)$$

$\because$ the additional term depends on $T(s)$, the relative value of states might change.

Example: This is because the states that are "closer" to terminal states will have low $T(s)$ which will increase $(\gamma)^{T(s)}$ which will decrease the additional term.

$V'_\pi(s)$ will have small value of additional term for states that are closer to the terminal states. It might change relative valuation of states.

Q5) Equation for v∗ in terms of q∗

Q5 Equation for v_* in terms of q_*.

we know, $v_*(s) = \max_{a \in A(s)} \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$.

and $q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a' \in A(s')} q_*(S', a') \mid S_t = s, A_t = a\right]$.

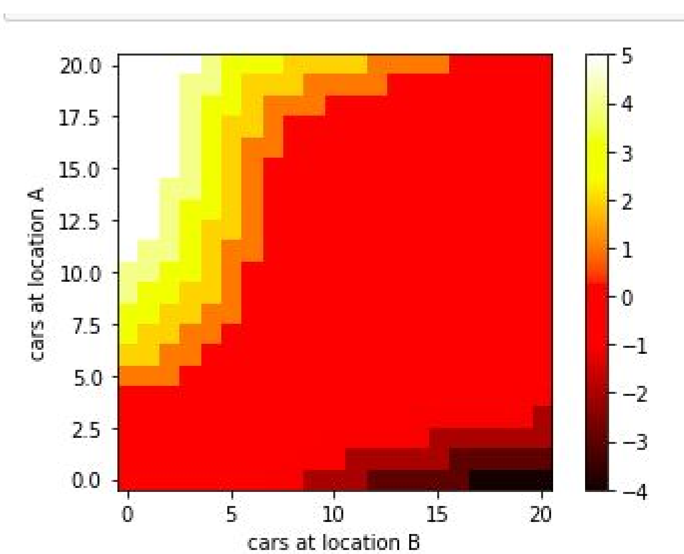
$\therefore \boxed{v_*(s) = \max_{a \in A(s)} q_*(s,a)}$.

Q6)

- policy iteration and value iteration (VI) to solve the Gridworld in Example 4.1
- the fix to the bug mentioned in Exercise 4.4. (in jupyter notebook)

Q7) Exercise 4.7

Original Example



Exercise - 4.7