Assignment 1 – Q3

Majid Ghasemi

September 2024

1 Q3

The general form of Bellman Equation for the state-value function is:

$$v_\pi(s) = \mathbb{E}_\pi \{G_t \mid S_t = s\} \tag{1}$$

Which can be written as:

$$\mathbb{E}_\pi \{R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\} \tag{2}$$

or without the expectation operator as:

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \left[\sum_{s' \in S} \left(\sum_{r \in R} p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right] \right) \right] \tag{3}$$

In this question, I will use Eq. 4 as another form of Bellman Equation to answer the question.

$$v_\pi(s) = \sum_{a \in A} \pi(a \mid s) \left[\sum_{s' \in S} p(s' \mid s, a) \left[r + \gamma v_\pi(s') \right] \right] \tag{4}$$

1.1 Question (a)

As states in the question, we have the equiprobable policy, and the environment is deterministic, in which choosing actions are taking place deterministically. γ is 0.9. Here is a list of notations used for this question:

- N : North
- S : South
- W : West
- E : East

We can define the reward function as follows as well:

$$r(s, a, s') \tag{5}$$

Where s is the current state, a is the taken action, and s' is the next state that the agent ends up there after taking action a.

Now, we can rewrite the Bellman Equation without any summations.

$$\begin{aligned} v_\pi(s_5) = \frac{1}{4} \left[r(s_5, N, s_0) + \gamma v_\pi(s_0) \right] + \frac{1}{4} \left[r(s_5, S, s_{10}) + \gamma v_\pi(s_{10}) \right] \\ + \frac{1}{4} \left[r(s_5, W, s_5) + \gamma v_\pi(s_5) \right] + \frac{1}{4} \left[r(s_5, E, s_6) + \gamma v_\pi(s_6) \right] \end{aligned} \tag{6}$$

Only the reward of hitting the wall and staying in s_5 (action W) is not zero (it is -1). Hence, we can rewrite it as follows:

$$\begin{aligned} v_\pi(s_5) = \frac{1}{4}\left[0+\gamma v_\pi(s_0)\right] + \frac{1}{4}\left[0+\gamma v_\pi(s_{10})\right] \\ + \frac{1}{4}\left[-1+\gamma v_\pi(s_5)\right] + \frac{1}{4}\left[0+\gamma v_\pi(s_6)\right] \end{aligned} \tag{7}$$

Substituting $\gamma = 0.9$, we have:

$$\begin{aligned} v_\pi(s_5) = \frac{1}{4}\left[0+0.9 v_\pi(s_0)\right] + \frac{1}{4}\left[0+0.9 v_\pi(s_{10})\right] \\ + \frac{1}{4}\left[-1+0.9 v_\pi(s_5)\right] + \frac{1}{4}\left[0+0.9 v_\pi(s_6)\right] \end{aligned} \tag{8}$$

We can simplify it further:

$$v_\pi(s_5) = \frac{1}{4}\left[\left[0+0.9 v_\pi(s_0)\right] + \left[0+0.9 v_\pi(s_{10})\right] + \left[-1+0.9 v_\pi(s_5)\right] + \left[0+0.9 v_\pi(s_6)\right]\right] \tag{9}$$

We can even simplify it more to be more clean and neat looking equation:

$$v_\pi(s_5) = \frac{1}{4}\left[0.9\left[\left(v_\pi(s_0)+v_\pi(s_{10})+v_\pi(s_6)\right) + \left(-1+v_\pi(s_5)\right)\right]\right] \tag{10}$$

I factored the gamma and put the states without reward in one place, and s_5 in another parenthesis.

Similarly, for s_{12}, we can have this series of equation. However, to save time for the readers, I will just give the final equation. All the rewards are zero, and by taking action N, agent goes to s_6, by action S goes to s_{17}, action W takes it to s_{11}, and action E takes it to be in s_{13}:

$$v_\pi(s_{12}) = \frac{1}{4}\left[0.9\left[\left(v_\pi(s_6)+v_\pi(s_{17})+v_\pi(s_{11})+v_\pi(s_{13})\right)\right]\right] \tag{11}$$

1.2 Question (b)

Using Eq. 6, we can answer this question. However, it must be noted that the value of $\pi(a|s)$ is different than the previous policy, and is given in the question.

Rewriting the Eq. 6, we have:

$$\begin{aligned} v_\pi(s_5) = 0.2\left[r(s_5,N,s_0)+\gamma v_\pi(s_0)\right] + 0.2\left[r(s_5,S,s_{10})+\gamma v_\pi(s_{10})\right] \\ + 0.1\left[r(s_5,W,s_5)+\gamma v_\pi(s_5)\right] + 0.5\left[r(s_5,E,s_6)+\gamma v_\pi(s_6)\right] \end{aligned} \tag{12}$$

Again, only the reward of hitting the wall is not zero, and we can rewrite the equation above as:

$$\begin{aligned} v_\pi(s_5) = 0.2\left[0+\gamma v_\pi(s_0)\right] + 0.2\left[0+\gamma v_\pi(s_{10})\right] \\ + 0.1\left[-1+\gamma v_\pi(s_5)\right] + 0.5\left[0+\gamma v_\pi(s_6)\right] \end{aligned} \tag{13}$$

By substituting the numbers from the given table as v values, and gamma, we have:

$$\begin{aligned} v_\pi(s_5) = 0.2\left[0+0.9\times 3.3\right] + 0.2\left[0+0.9\times 0.1\right] \\ + 0.1\left[-1+0.9\times 1.5\right] + 0.5\left[0+0.9\times 3\right] \end{aligned} \tag{14}$$

Is equal to: 1.997.

1.3 Question (c)

Similar to question (b), for $v_\pi(s_{12})$, after simplifications, we have:

$$\begin{aligned} v_\pi(s_{12}) = 0.6\left[0+0.9\times 2.3\right] + 0.02\left[0+0.9\times -0.4\right] \\ + 0.08\left[0+0.9\times 0.4\right] + 0.3\left[0+0.9\times 0.7\right] \end{aligned} \tag{15}$$

Where it is equal to: 1.4526.

1.4 Question (d)

Based on the achieved results in the previous sections, we can see that the results under different policies, outperformed our initial equiprobable policy.

SIGNED AS: MAJID GHASEMI , `September 22`