AI 3000 : Reinforcement Learning
Assignment No. 1

Ahmik Virani
ES22BTECH11001

1. Markov Reward Process
   (a) Identify the states, transition matrix of this Markov process.
   (b) Construct a suitable reward function, discount factor and use the Bellman equation for the Markov reward process to compute how long does it take "on average" (the expected number of dice throws) to reach the state W from any other state.

Ans: A1.

(a) The states are all the 'squares' that can be occupied.

$$S = \{\text{S}, 1, 2, 3, 4, 5, 6, 7, 8, 9, \text{W}\}$$

The transisition matrix will be the probability of going from one state to any other state, thus it would be a 11x11 matrix.

$P =$

	S	1	2	3	4	5	6	7	8	9	W
S	0	$\frac{1}{4}$	0	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$	0	0
1	0	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}$	0	0
2	0	0	0	0	0	0	0	1	0	0	0
3	0	0	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0	0	0
4	0	0	0	0	0	0	0	0	1	0	0
5	0	0	0	$\frac{1}{4}$	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0
6	0	0	0	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{4}$
7	0	0	0	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{4}$
8	0	0	0	$\frac{1}{4}$	0	0	0	0	$\frac{1}{2}$	0	$\frac{1}{4}$
9	0	0	0	1	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	1

(b) We want to compute how long does it take on average to reach state W from any other state. Thus we can construct the following reward function: Reward R: -1 for every move made until reaching goal state(W). More formally,

$$R(s) = -1, \text{for s} \in \{\text{S}, 1, 3, 5, 6, 7, 8, \text{W}\}$$

$$R(s) = 0, \text{otherwise}$$

The reason for choosing -1 as reward for most of the states is that it indicates the number of moves you take. Every extra move you take, the value of cumulative reward decreases by 1, and thus it counts the number of steps taken.

The reason for reward for states 2, 4 and 9 being 0 is that, where there is a snake/ladder, it is equivalent to not taking any move and going to where the snake/ladder directs. Thus the reward need not add an extra step, because we are essentially just teleporting and are never going to roll a dice at the position of exactly 2, 4, or 9. Thus being at 2 is equivalent to being at 7, and so on.

Similarly, at state W, you are at the terminal state, which is the absorbing state. Thus you are done with the game and do not need any more moves.

Since this is a finite horizon problem, we can set the discount factor, $\gamma = 1$

Next, we can use the Bellman equation in matrix form to compute the value function at each state: (Please check this Google Colab notebook to see the python code for the same).

$$\begin{aligned}
V(s_{\mathrm{S}}) &= -7.0833 \\
V(s_1) &= -7.0000 \\
V(s_2) &= -5.3333 \\
V(s_3) &= -6.6667 \\
V(s_4) &= -5.3333 \\
V(s_5) &= -6.6667 \\
V(s_6) &= -5.3333 \\
V(s_7) &= -5.3333 \\
V(s_8) &= -5.3333 \\
V(s_9) &= -6.6667 \\
V(s_{\mathrm{W}}) &= 0.0000
\end{aligned}$$

Thus the expected number of steps to reach from a particular state to W is

(Note $\mathbb{E}(s_i)$ denotes the expected number of steps from state i to reach W):

$$\begin{aligned}
\mathbb{E}(s_{\mathrm{S}}) &= 7.0833 \\
\mathbb{E}(s_1) &= 7.0000 \\
\mathbb{E}(s_2) &= 5.3333 \\
\mathbb{E}(s_3) &= 6.6667 \\
\mathbb{E}(s_4) &= 5.3333 \\
\mathbb{E}(s_5) &= 6.6667 \\
\mathbb{E}(s_6) &= 5.3333 \\
\mathbb{E}(s_7) &= 5.3333 \\
\mathbb{E}(s_8) &= 5.3333 \\
\mathbb{E}(s_9) &= 6.6667 \\
\mathbb{E}(s_{\mathrm{W}}) &= 0.0000
\end{aligned}$$

2. Effect of Noise and Discounting

   (a) Identify what values of γ and η lead to each of the optimal paths listed above with reasoning. If necessary, you could implement the value iteration algorithm on this environment and observe the optimal paths for various choices of γ and η.

Ans:

(a) We are given

- S = the set of squares shown in the figure that could be visited.
- A = either go left, right, up or down at each state. If an action takes the agent off the grid, the state remains unchanged.
- $R = \begin{cases} +1 & \text{if in blue state,} \\ +10 & \text{if in green state,} \\ -10 & \text{if in red state,} \end{cases}$

  the red, green, blue states mentioned are from Figure 2. given in the problem statement.

To build the basic intuition, let us say that $\gamma \approx 0$, then we are mostly bothered about the immediate reward (i.e. myopic). Whereas if a higher value of γ would mean that we go towards far-sightedness. Thus a lower

value of γ would mean that we fine with the nearby reward, whereas a larger γ would mean that we expect a higher reward by going far.

However, if we bring the noise into picture, if the noise η is very high, then we would be more inclined in going for a more safer path than take a path which has high risk, because the value function would be affected by a significant probability of taking a 'red' state and getting reward -10 in the riskier (dashed line) path due to the highly random nature. If η is low, then we could go in the more riskier path, as the low probability of going into the red square would bring down the effect of the high negative reward making it not very significant.

Thus to put the above together:

- Low γ, Low η: Prefer the close exit (state with reward +1) but risk the cliff (dashed path to +1)
- High γ, Low η: Prefer the distant exit (state with reward +10) but risk the cliff (dashed path to +10)
- Low γ, High η: Prefer the close exit (state with reward +1) by avoiding the cliff (solid path to +1)
- High γ, High η: Prefer the distant exit (state with reward +10) by avoiding the cliff (solid path to +10)

3. Markov Decision Process

Ans:

(a) From the State Value Function definition, we get the expression for $V^\pi(s)$ and $\hat{V}^\pi(s)$ as follows:

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s\right]$$

$$\hat{V}^\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \hat{r}_{t+k+1} \middle| s_t = s\right]$$

where

$$r_{t+k+1} = R(s_{t+k}, a_{t+k}, s_{t+k+1})$$

and

$$\hat{r}_{t+k+1} = \hat{R}(s_{t+k}, a_{t+k}, s_{t+k+1})$$

Subtracting the above two equations and using the fact that $R(s,a,s') - \hat{R}(s,a,s') = \epsilon$, we get:

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \epsilon \middle| s_t = s\right]$$

Since ϵ and γ are constants we can pull that out of the expectation:

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \sum_{k=0}^{\infty} \gamma^k \epsilon$$

As $\gamma < 1$, we can write the above as:

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \frac{\epsilon}{1-\gamma}$$

Rearranging, we get:

$$V^{\pi}(s) = \hat{V}^{\pi}(s) + \frac{\epsilon}{1-\gamma}$$

(b) Yes, they have the same optimal policy.

Let us say that π^* is the optimal policy for M. This means that:

$$V^{\pi^*}(s) \geq V^{\pi}(s), \forall s \in S$$

Subtracting $\frac{\epsilon}{1-\gamma}$ from both sides

$$\hat{V}^{\pi^*}(s) \geq \hat{V}^{\pi}(s), \forall s \in S$$

Thus, π^* is an optimal policy for $\hat{M}$ too.

Conclusion, if value function differs by a constant, then both the MPD's have the same optimal policies.

(c) As it is finite horizon, let us specify the time units as $t = 0, 1, ..., H-1$, where H is horizon length.

Following similar steps as in part(a) of this question, we have:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{H-1} \gamma^k \; r_{t+k+1} \middle| s_t = s\right],$$

$$\hat{V}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{H-1} \gamma^k \, \hat{r}_{t+k+1} \,\middle|\, s_t = s\right].$$

Subtracting the above two equations gives:

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{H-1} \gamma^k (r_{t+k+1} - \hat{r}_{t+k+1}) \middle| s_t = s\right].$$

Using $R(s,a,s') - \hat{R}(s,a,s') = \epsilon$, we get

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{H-1} \gamma^k \epsilon \middle| s_t = s\right]$$

Since both terms inside the expectation are constants, we get:

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \sum_{k=0}^{H-1} \gamma^k \epsilon$$

Thus,

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \epsilon \frac{(1-\gamma^H)}{(1-\gamma)}$$

Rearranging: Thus,

$$V^{\pi}(s) = \hat{V}^{\pi}(s) + \epsilon \frac{(1-\gamma^H)}{(1-\gamma)}$$

(d) In finite horizon we had a definite value of H allowing us to remove the expectation. However, in indefinite horizon, since the number of steps, H as defined in the question, is itself random, thus we cannot take away the expectation. Therefore, the equation remains:

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{H-1} \gamma^k \, \varepsilon \,\middle|\, s_t = s\right].$$

Or simplifying further:

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \frac{\epsilon}{1-\gamma}\left(1 - \mathbb{E}_{\pi}\left[\gamma^H \,\middle|\, s_t = s\right]\right).$$

Thus the analogous result of part (a) does not hold exactly since we require an extra expectation due to the uncertainity of the length of the trajectory. This is dependent on the policy as well as the current state, which is different from part (a).

(e) From part (a) I reiterate the formula for $V^{\pi}(s)$ and $\hat{V}^{\pi}(s)$:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s\right]$$

$$\hat{V}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \hat{r}_{t+k+1} \middle| s_t = s\right]$$

Subtracting the above two equations:

$$V^{\pi}(s) - \hat{V}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \left(r_{t+k+1} - \hat{r}_{t+k+1}\right) \middle| s_t = s\right]$$

Thus, using the fact that $\left|R(s,a,s') - \hat{R}(s,a,s')\right| \leq \epsilon$, we get:

$$\left|V^{\pi}(s) - \hat{V}^{\pi}(s)\right| \leq \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \epsilon \middle| s_t = s\right] = \frac{\epsilon}{1-\gamma}$$

Now, let π^* be the optimal policy for M and let $\hat{\pi}^*$ be the optimal policy for $\hat{M}$.

Thus, let us write,

$$\left|V^{\pi^*}(s) - \hat{V}^{\pi^*}(s)\right| \leq \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \epsilon \;\middle|\; s_t = s\right] = \frac{\epsilon}{1-\gamma} \tag{1}$$

$$\left|V^{\hat{\pi}^*}(s) - \hat{V}^{\hat{\pi}^*}(s)\right| \leq \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \epsilon \middle| s_t = s\right] = \frac{\epsilon}{1-\gamma} \tag{2}$$

We also know that:

$$\hat{V}^{\hat{\pi}^*}(s) \geq \hat{V}^{\pi^*}(s)$$

$$V^{\pi^*}(s) \geq V^{\hat{\pi}^*}(s)$$

Let us take 2 cases:
(i)$\hat{V}^{\hat{\pi}^*}(s) \geq V^{\pi^*}(s)$ this is equivalent to $\hat{V}_*(s) \geq V_*(s)$
(ii)$\hat{V}^{\hat{\pi}^*}(s) \leq V^{\pi^*}(s)$ this is equivalent to $\hat{V}_*(s) \leq V_*(s)$
First let us solve case (i):

We have

$$\hat{V}_* \geq V_* \geq V^{\hat{\pi}^*} \tag{3}$$

Thus 2 would open as:

$$\hat{V}^{\hat{\pi}^*}(s) - V^{\hat{\pi}^*}(s) \leq \frac{\epsilon}{1-\gamma}$$

From 3, we have

$$\hat{V}^{\hat{\pi}^*}(s) - V^{\pi^*}(s) \leq \hat{V}^{\hat{\pi}^*}(s) - V^{\hat{\pi}^*}(s)$$

and hence, under condition that $\hat{V}_*(s) \geq V_*(s)$, we get:

$$\hat{V}^{\hat{\pi}^*}(s) - V^{\pi^*}(s) \leq \frac{\epsilon}{1-\gamma}$$

Next let us solve case (ii):

We have

$$V_*(s) \geq \hat{V}_*(s) \geq \hat{V}^{\pi^*}(s) \tag{4}$$

Thus 1 would open as:

$$V^{\pi^*}(s) - \hat{V}^{\pi^*}(s) \leq \frac{\epsilon}{1-\gamma}$$

From 4, we have

$$V^{\pi^*}(s) - \hat{V}^{\hat{\pi}^*}(s) \leq V^{\pi^*}(s) - \hat{V}^{\pi^*}(s)$$

and hence, under condition that $V_*(s) \geq \hat{V}_*(s)$, we get:

$$V^{\pi^*}(s) - \hat{V}^{\hat{\pi}^*}(s) \leq \frac{\epsilon}{1-\gamma}$$

Thus, we have gotten:

- From (i): $\hat{V}_*(s) - V_*(s) \leq \frac{\epsilon}{1-\gamma}$ if $\hat{V}_*(s) \geq V_*(s)$
- From (ii) : $V_*(s) - \hat{V}_*(s) \leq \frac{\epsilon}{1-\gamma}$ if $V_*(s) \geq \hat{V}_*(s)$

Hence, we conclude

$$\left| V_*(s) - \hat{V}_*(s) \right| \leq \frac{\epsilon}{1-\gamma}$$

From the above derivation, we can see that for the given criteria, the optimal policy for M, which is π^* and the optimal policy for $\hat{M}$, which is $\hat{\pi}^*$, need not be same. Infact, we got the inequalty because of the fact that the two are not same. In the above derived relation between $V_*(s)$ and $\hat{V}_*(s)$, the equality would hold only when the optimal policy are same (as showed in part (a)). Every time else, when there is a strict inequality, the two optimal policies need not be the same.

The same idea can be thought intuitively as follows: Since we get different reward in both MDP's, and they differ atmost by a constant ϵ, the reward for going from s to s' by taking action a in M could be more than that in $\hat{M}$. For $\hat{M}$, maybe going from s to s'' by taking action a' gives a more reward, and the value function $V(s')$ be same as $\hat{V}(s'')$, thus it would be beneficial to take action a in M but a' in $\hat{M}$. Hence by proof by construction, I have showed that the optimal policy need not be the same.