Assignment 1
Building Reinforcement Learning Environment
Report

Charanteja Kalva
UBIT : charante
charante@buffalo.edu

1 Deterministic/Stochastic Environments

1.1 Deterministic Environment defined:

In this assignment I have developed a square grid world with following features

- The State set of the Deterministic environment created involved the below states : total 16 in count

$S_0 - (0,0)$	$S_1 - (0,1)$	$S_2 - (0,2)$	$S_3 - (0,3)$
$S_4 - (1,0)$	$S_5 - (1,1)$	$S_6 - (1,2)$	$S_7 - (1,3)$
$S_8 - (2,0)$	$S_9 - (2,1)$	$S_{10} - (2,2)$	$S_{11} - (2,3)$
$S_{12} - (3,0)$	$S_{13} - (3,1)$	$S_{14} - (3,2)$	$S_{15} - (3,3)$

Table 1: States of the Environment

- The subscript followed by S gives the position of the state in the Observation list which I used interchangeably to identify the state of agent in the implementation.
- In my implementation I have chosen to have **4** actions i.e. **down, up, right and left**. **4** rewards which are **0, 2, 3, 5**. Here reward **5** is for the goal state.
- Action set = A = **{down, up, right, left}**
- Reward set = R = **{0, 2, 3, 5 }**
- The main objective is to reach the goal with maximum cumulative reward.

1.2 Stochastic Environment defined:

- The Stochastic Environment was a bit modified version of the Deterministic environment which had an extra feature of applying probability(uncertainty) for the transition from current state to next state after certain action is performed.
- Here in this scenario, agent after taking an action it has a probability of **0.8** to move to the respective following state and with probability **0.2** it can stay in the current state.
- The state set is same as Deterministic environment with 16 states as below.

$S_0 - (0,0)$	$S_1 - (0,1)$	$S_2 - (0,2)$	$S_3 - (0,3)$
$S_4 - (1,0)$	$S_5 - (1,1)$	$S_6 - (1,2)$	$S_7 - (1,3)$
$S_8 - (2,0)$	$S_9 - (2,1)$	$S_{10} - (2,2)$	$S_{11} - (2,3)$
$S_{12} - (3,0)$	$S_{13} - (3,1)$	$S_{14} - (3,2)$	$S_{15} - (3,3)$

Table 2: States of the Environment

- Action set in the Stochastic Environment has an additional action called **stay**.
- Action Set = A = **{down, up, right, left, stay}**
- The reward set also has an additional reward for staying in the current state even after action is performed i.e **-1**
- Reward set = R = **{-1, 0, 2, 3, 5 }**
- The main objective is to reach the goal with maximum cumulative reward.

2 Deterministic/Stochastic Environments Differences

2.1 Deterministic Environment:

- In this environment if an action is chosen by the agent there is no uncertainty in the outcome of that action.
- In other words, after an action action is chosen by the agent and if there is a possible next state for that action, the agent should end up in that state with probability 1 (i.e. compulsorily)

2.2 Stochastic Environment:

- In this environment if an action is chosen by the agent there will an uncertainty associated with the output of that action.
- After an action has been chosen to be performed by an agent, there might be multiple possible states to be reached from current state using the given action and all those possible states are associated with some probabilities, this makes the environment Stochastic.
- For example, if the agent is in state S_1 and take action a_1 and this leads to two states S_2 and S_3 with probabilities 0.6 and 0.4 respectively then this scenario becomes a stochastic environment.
- In terms of assignment, I have chosen my Stochastic environment in such a way that the agent follows the action 80% of the time and remaining 20% it stays in the current state, this is what makes it stochastic.

3 Transition-probability matrix for Stochastic Environment.

Consider the following table for referring the positions of the Grid world to the states.

S_0	S_1	S_2	S_3
S_4	S_5	S_6	S_7
S_8	S_9	S_{10}	S_{11}
S_{12}	S_{13}	S_{14}	S_{15}

Table 3: States of the Environment

- Following Figure 1 is the transition table which shows the outcomes of the actions for every state and every action taken.
- Table 4 in the next page describe the probabilities that apply for transitions from one state to other state in the stochastic environment.

. **state - [down, up, right, left]**

```
S0 - [4. 0. 1. 0.]
S1 - [5. 1. 2. 0.]
S2 - [6. 2. 3. 1.]
S3 - [7. 3. 3. 2.]
S4 - [8. 0. 5. 4.]
S5 - [9. 1. 6. 4.]
S6 - [10.  2.  7.  5.]
S7 - [11.  3.  7.  6.]
S8 - [12.  4.  9.  8.]
S9 - [13.  5. 10.  8.]
S10 - [14.  6. 11.  9.]
S11 - [15.  7. 11. 10.]
S12 - [12.  8. 13. 12.]
S13 - [13.  9. 14. 12.]
S14 - [14. 10. 15. 13.]
S15 - [15. 11. 15. 14.]
```

Figure 1: Transitions of each state

	S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}
S_0	0	**0.8**	0	0	**0.8**	0	0	0	0	0	0	0	0	0	0	0
S_1	**0.8**	0	**0.8**	0	0	**0.8**	0	0	0	0	0	0	0	0	0	0
S_2	0	**0.8**	0	**0.8**	0	0	**0.8**	0	0	0	0	0	0	0	0	0
S_3	0	0	**0.8**	0	0	0	0	**0.8**	0	0	0	0	0	0	0	0
S_4	**0.8**	0	0	0	0	**0.8**	0	0	**0.8**	0	0	0	0	0	0	0
S_5	0	**0.8**	0	0	**0.8**	0	**0.8**	0	0	**0.8**	0	0	0	0	0	0
S_6	0	0	**0.8**	0	0	**0.8**	0	**0.8**	0	0	**0.8**	0	0	0	0	0
S_7	0	0	0	**0.8**	0	0	**0.8**	0	0	0	0	**0.8**	0	0	0	0
S_8	0	0	0	0	**0.8**	0	0	0	0	**0.8**	0	0	**0.8**	0	0	0
S_9	0	0	0	0	0	**0.8**	0	0	**0.8**	0	**0.8**	0	0	**0.8**	0	0
S_{10}	0	0	0	0	0	0	**0.8**	0	0	**0.8**	0	**0.8**	0	0	**0.8**	0
S_{11}	0	0	0	0	0	0	0	**0.8**	0	0	**0.8**	0	0	0	0	**0.8**
S_{12}	0	0	0	0	0	0	0	0	**0.8**	0	0	0	0	**0.8**	0	0
S_{13}	0	0	0	0	0	0	0	0	0	**0.8**	0	0	**0.8**	0	**0.8**	0
S_{14}	0	0	0	0	0	0	0	0	0	0	**0.8**	0	0	**0.8**	0	**0.8**
S_{15}	0	0	0	0	0	0	0	0	0	0	0	**0.8**	0	0	**0.8**	0

Table 4: Transition Probabilities between states

4 Main Components of RL Environment:

- Generally RL environment should have State set, Action set, Reward set, gamma, transition probability matrix, policy, value function as the parameters to define the environment.
- State Set is the set of all the states that are present in the environment. All possible combinations of where an agent can end up in the environment.
- Action set is the set of all possible action an agent can perform. Once an action is performed by an agent, it is the environment that decides the possible resulting state based on the probabilities. If the probability is 0 or 1 then its is deterministic else stochastic environment.
- Reward Set is the set of all possible rewards an agent can gain in the environment.
- Gamma is the discount factor which reduces the impact of future reward. It can also control the direction of agent in order to reach the goal.
- Transition Probability Matrix defines the probabilities of all possible state transitions. After an action is performed the probability of new state will depend on this matrix
- Policy determines what action an agent should take if it is present in a particular state. In brief policy determines the behaviour of an agent.
- Value Function gives the value of states that enables us to judge the action we need to take in order to reach the goal or reward.
- Below is the value function which was used to calculate the state value of every state. Once the values are converged, based on the values of the states we obtain the policy which the agent should follow to reach the goal with maximum cumulative reward.

$$\mathrm{V_{k+1}(S)} = \sum_{\mathrm{a}} \pi(a|s) \sum_{s',a} p(s', r|s, a)[r + \gamma V_k(S')]$$

5 Observation of Agent Movements in the Environments:

5.1 In Deterministic Environment :

- In Deterministic Environment, once after getting the state values of all the states, the agent uses those values to find the optimal policy of the state in which the agent is present and take action according to that optimal policy.
- In 6 steps the agent reaches the goal.

- Observed changing different values of gamma, for some values of gamma the agent goes into loop between states and for some values of gamma the agent successfully reaches goal. The same scenario happens in case of stochastic environment also.

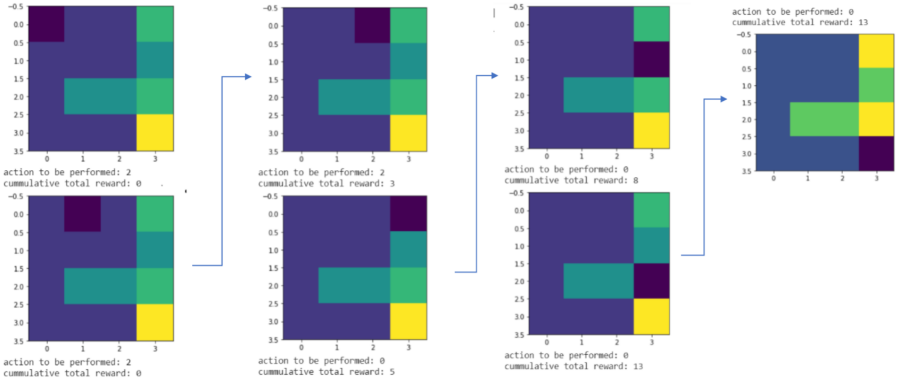

Figure 2: Movement of agent in Deterministic Environment

5.2 In Stochastic Environment :

- In Stochastic Environment, once after getting the state values of all the states, the agent uses those values to find the optimal policy of the state in which the agent is present and take action according to that optimal policy.
- Now the action is sent to environment to update the state but as it is Stochastic, the change of state depends on probability i.e. sometimes it gets updated and sometimes it does not update.

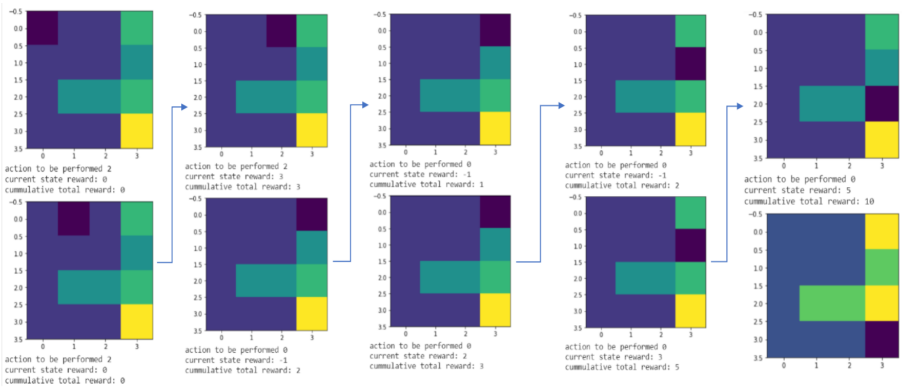

Figure 3: Movement of agent in Stochastic Environment

- You can observe this when the agent is in right top corner, the action to be performed is 0 i.e. down but it stayed in the current state because of stochastic nature.
- In this the agent took 9 steps to reach the goal.

6 Tabular Method used:

- In this assignment I have used Dynamic Programming as the tabular method to achieve the main objective.
- In DP there are two phases
  - Policy Evaluation
  - Policy Improvement
- We start with a random policy and evaluate the policy by calculating the state values using that policy.
- Next we improve the policy by observing the values of the transition states($s^{'}$) of a particular state(s). We update our policy in such a way that the agent takes actions in the direction of maximum state value.

$$\mathrm{V_{k+1}(S)} = \sum_{\mathrm{a}} \pi(a|s) \sum_{s^{'},a} p(s^{'}, r|s, a)[r + \gamma V_k(S^{'})]$$

- I have used the above formula (random policy) for calculating initial state values of all the states present i.e. 16 states.
- Finally, agent takes the action depending on the value of neighbouring states. Thus the agent will achieve maximum cumulative reward.

7 Implementation Details:

- My implementation is flexible to generate a Dynamic environment. That is a grid environment of variable size with randomly generated reward positions can be built.
- But to maintain consistency with report I have made the positions static. And I have mentioned needed changes to be done in the code to generate dynamic grid as comments.
- To check for other sizes, update the variables **tot_states, grid_size**
- To change number of rewards in the grid change the variable **tot_rewards**
- As mentioned, if agent falls in a loop between states, try changing gamma value to observe the change in policy of agent.
- If grid size is increased make sure **max_timesteps** is also increased