**Reinforcement Learning Assignment #1 Report**

**Hazique Quazi**

[mquazi@buffalo.edu](mailto:mquazi@buffalo.edu)

UB Person #: 50416875

[**Github Repo for Submissions**](https://github.com/hazique/CSE546-RL-Assignments-Spring-22) **(@ub-rl added as collaborator)**

**Abstract**

This report details and demonstrates our learning of Reinforcement Learning concepts such as creating an agent and environment, running the agent for a given number of timesteps and visualizing the change of states and rewards given to the agent by the environment in response to a given action. Below I describe the execution of the agent in Deterministic and Stochastic Environment.

# Deterministic Environment

The deterministic environment can be described as follows:

Action Set

State Set

Reward Set

The reward received by the agent can be better represented as a 4x4 matrix

([[0, 1, 1, 1],

[1, -1, -1, 1],

[1, -1, -1, 1],

[1, 1, 1, 2]])

The reward 0 corresponds to the starting position of the agent whereas the 2 corresponds to the end position.

### Visualization for Deterministic Environment

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart

Description automatically generated Chart

Description automatically generated Chart

Description automatically generated

Table

Description automatically generated

# Stochastic Environment

The Stochastic environment has the same Action Set, State Set & Reward Set as the Deterministic environment. It however adds stochasticity to the action the agent can take at any time step. I have defined the stochastic environment by letting an agent take an action probabilistically according to the following logic:

1. Assign random probabilities to actions (left = up = 0 to10%, right = down = 11-90%)
2. Check if direction with maximum probability has a probability value greater than 30%
   1. If False, the agent goes left instead of right OR the agent goes up instead of down.

### Visualization for Stochastic Environment

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart

Description automatically generated

Chart

Description automatically generated Chart, histogram

Description automatically generated Chart, bar chart, histogram

Description automatically generated

Table

Description automatically generated

# Training on Q Learn

### Deterministic Environment

Epsilon Decay – The epsilon value decays at a rate of 0.995 per episode for 1000 episodes



Total Rewards during Training



Total penalties during Training

### Inference:

As the number of penalties go down and the number of rewards collected saturates at a timesteps = 6, we can say that the agent has learned the optimal path.

Evaluation Results:

Since the agent earns constant reward in 6 timesteps while successfully reaching the goal position, our agent has in fact learned the optimal path.



### Stochastic Environment:

Epsilon Decay rate is 0.9977 per step for 2000 steps



The stochastic environment presents a challenge for the agent. As a result, the agent collects a wide range of rewards during the training cycle.



Concomitantly, the agent ends up taking a greater number of steps and penalties during the training.





Evaluation Results:

As a result of the environment being stochastic, the agent is not able to execute the optimal policy. It seldom reaches the goal position.



# Training on SARSA

### Deterministic Environment

Epsilon Decay – The epsilon value decays at a rate of 0.995 per episode for 1000 episodes



Total Rewards during Training



Total Penalties



### Inference:

As the number of penalties go down and the number of rewards collected saturates at minimum timesteps 6 or 7, we can say that the agent has learned the optimal path using SARSA algorithm.

Evaluation Result:

Since the agent earns constant reward in 6 timesteps while successfully reaching the goal position, our agent has in fact learned the optimal path.



### Stochastic Environment:

Epsilon Decay rate is 0.9977 per step for 2000 steps



The stochastic environment presents a challenge for the agent. As a result, the agent collects a wide range of rewards during the training cycle like Q Learn in a stochastic environment.



Similarly, the agent receives variable number of penalties throughout training.



Evaluation Results:

Similar to the performance in Q Learning, the agent seldom reaches the goal position.



# Q Learn vs SARSA Comparison

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Training | | | | Evaluation | | | |
|  | Q Learn Det | Q Learn Stoch | SARSA Det | SARSA Stoch | Q Learn Det | Q Learn Stoch | SARSA Det | SARSA Stoch |
| Avg steps | 7.661 | 18.672 | 7.981 | 18.754 | 6 | 18.5 | 6 | 18.9 |
| Total Reward | 26245 | -17000 | 24525 | -19470 | 300 | -35 | 300 | -125 |
| Successful | 920 | 410 | 915 | 388 | 10 | 2 | 10 | 3 |
| Penalties | 1206 | 2129 | 912 | 20701 | 0 | 96 | 0 | 107 |

From the above result we can clearly say that Q Learn outperform SARSA irrespective of the environment – deterministic or stochastic, on both training and evaluation routines.