CS6700: Reinforcement Learning
Programming Assignment 3

Shubham Patel(ME19B170), Sharuhasan(EE19B117)

1 Introduction

The assignment aims at solving the OpenAI Gym's Taxi-v3 environment.

RL agents are trained by **Hierarchical Reinforcement Learning** methodologies. Further, 2 different variations is used namely **SMDP** and **Intra Option** are tested. **Epsilon Greedy strategy** is employed to choose an action.

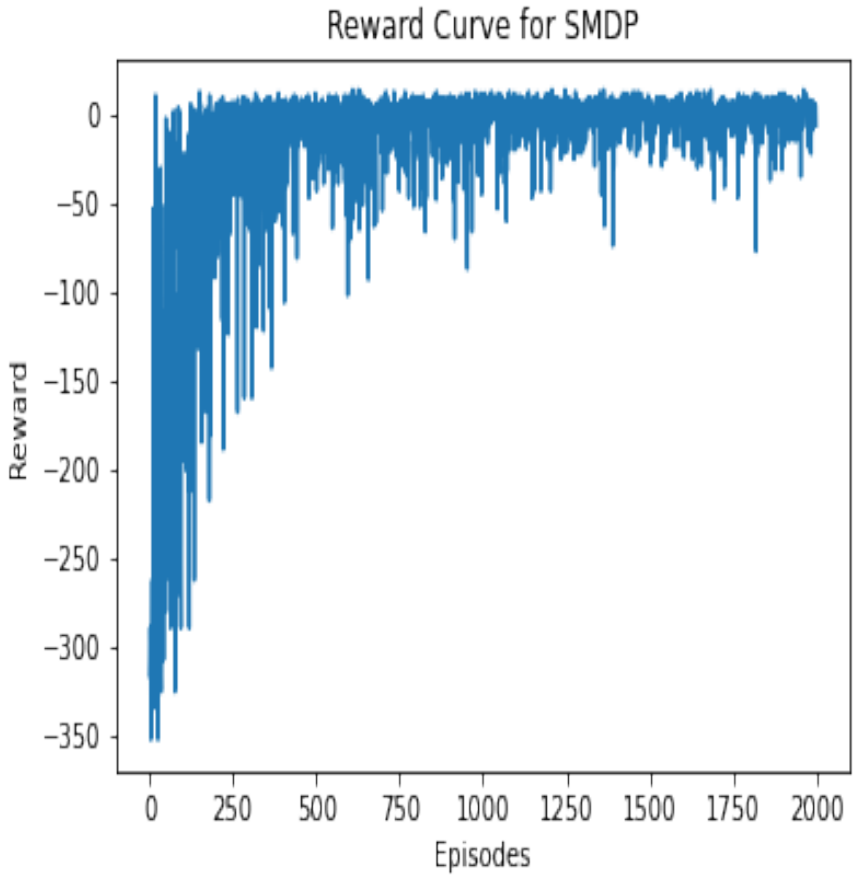
The environment provides action to move North, South, East, West, Pickup and Drop. Alongside 6 deterministic actions, four different options are added, for going to each of four designated locations, when the taxi is not already there.

2 SMDP

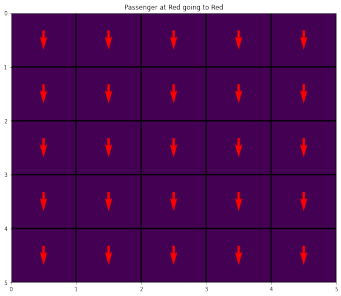
- **Hyper-parameters**
  - Gamma: 0.9
  - Alpha: 0.4
  - Episodes: 2000

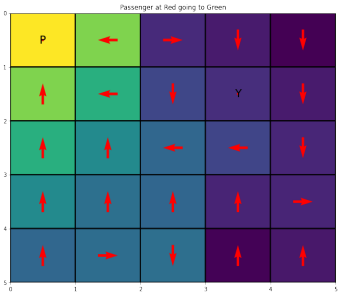
Results:

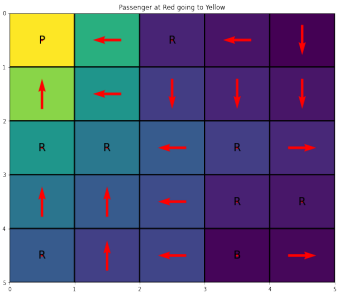
Average Reward for last 500 episodes = -0.304

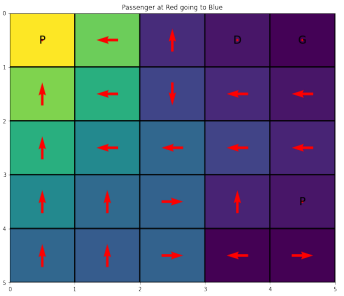


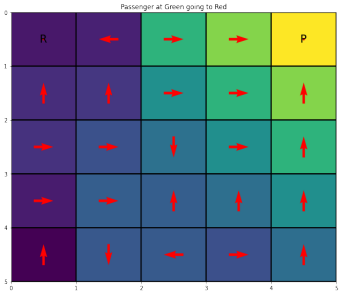
Visualisation of Q Table:

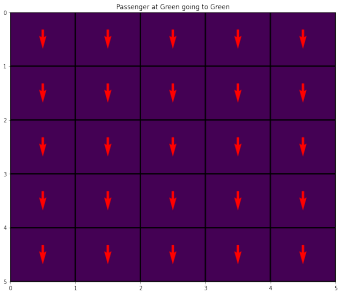


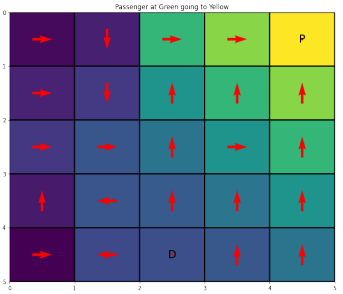


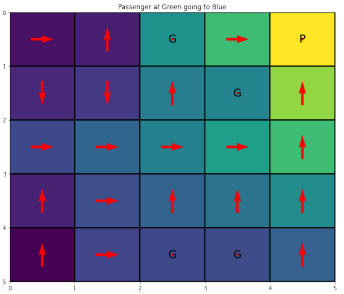


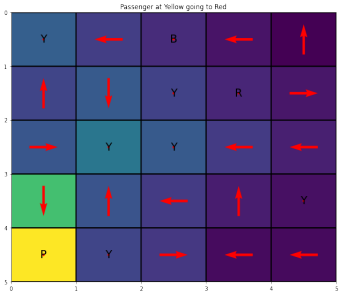


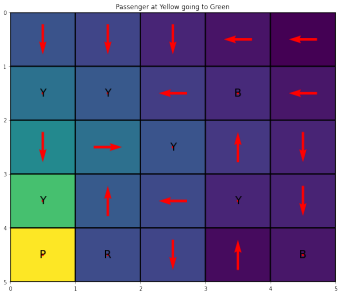


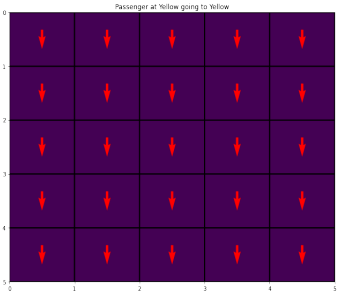


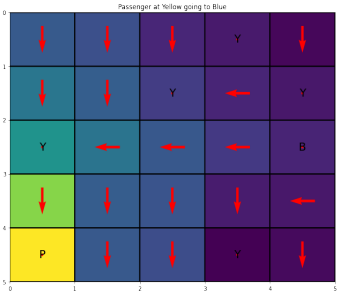


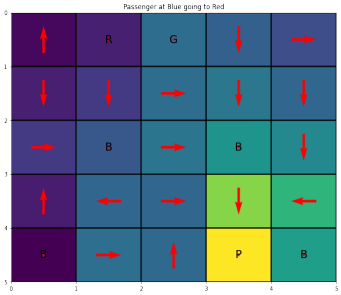


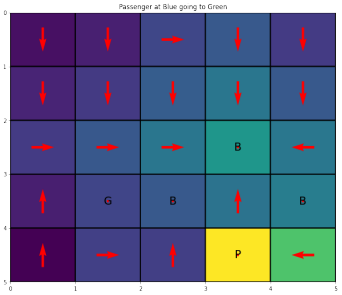


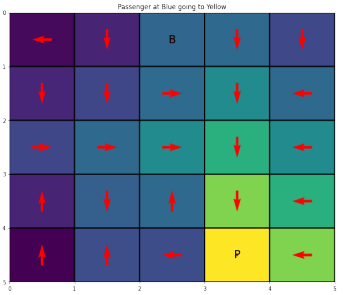


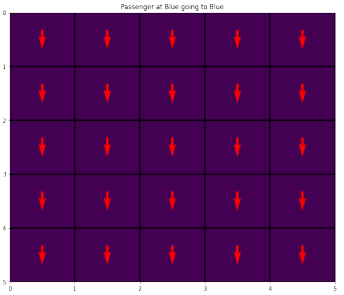


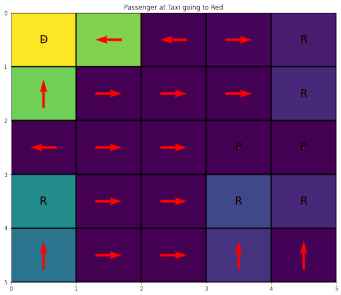


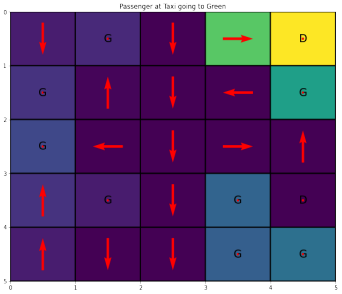


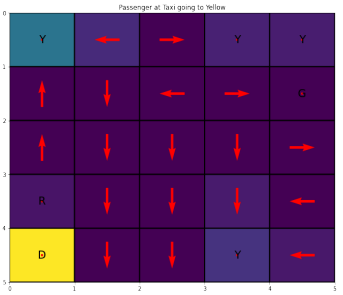


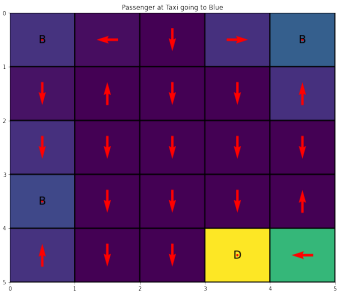










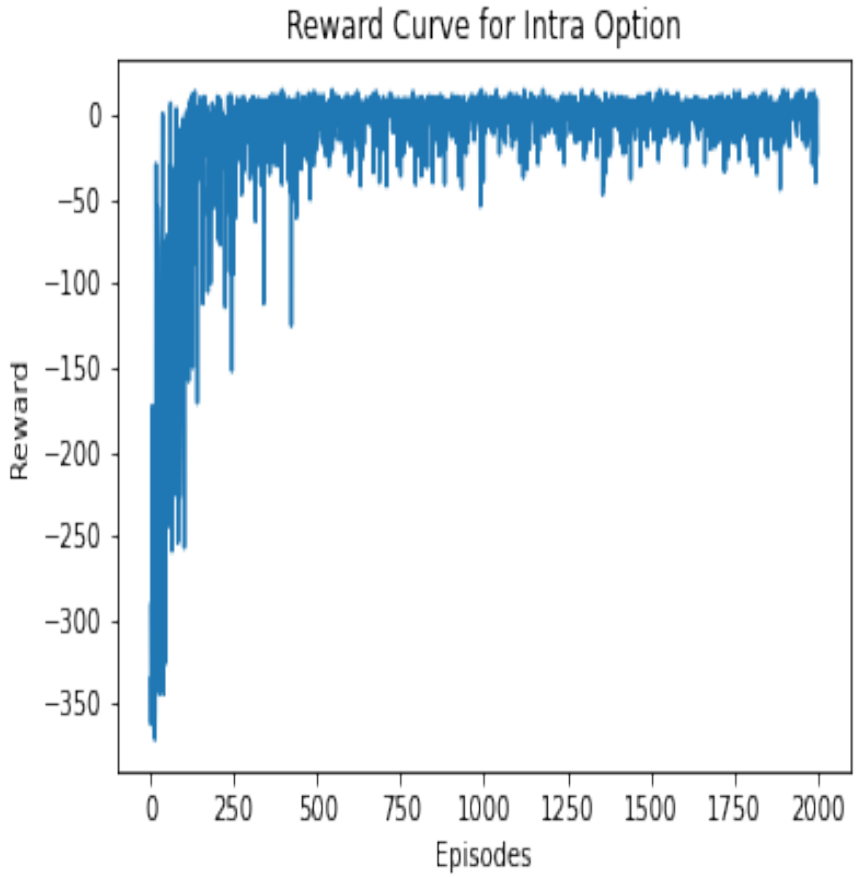


3 Intra Option

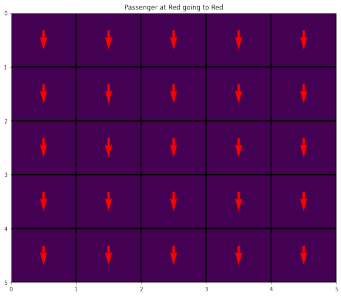
- **Hyper-parameters**
  - Gamma: 0.9
  - Alpha: 0.5
  - Episodes: 2000

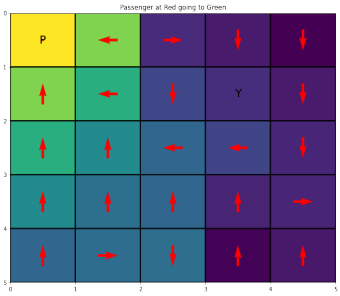
Results:

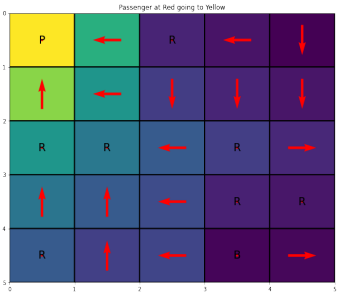
Average Reward for last 500 episodes = 0.64

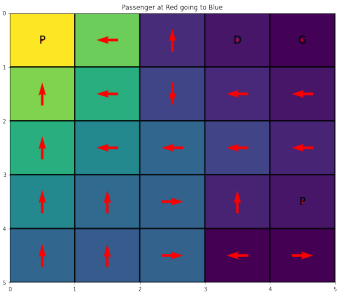


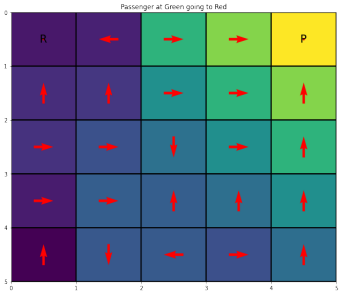
Visualisation of Q Table:

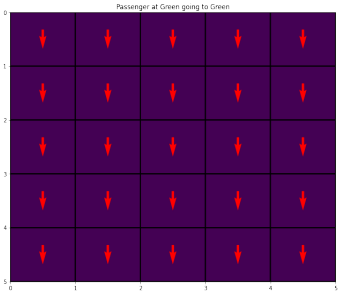


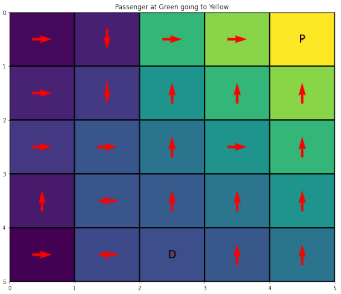


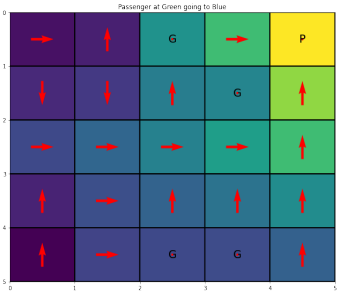


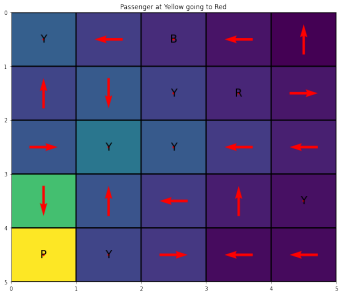


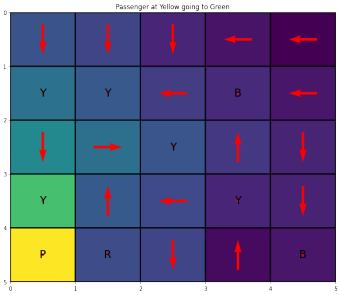


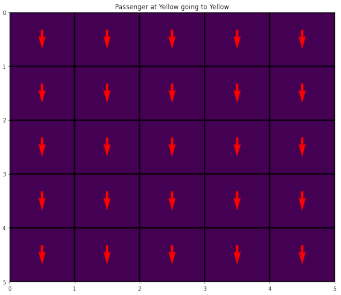


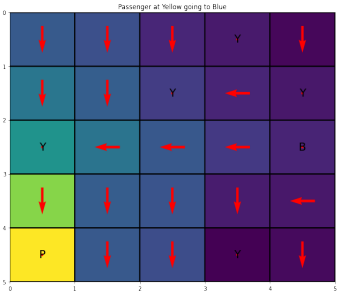


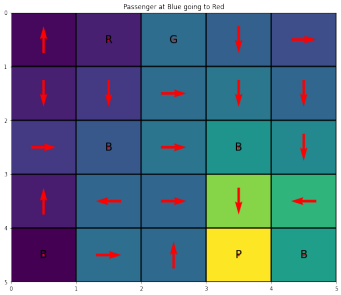


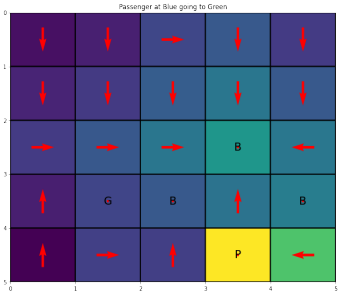


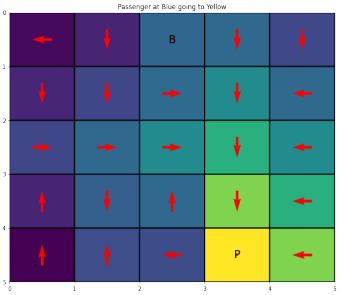


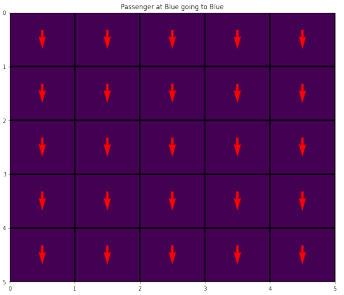


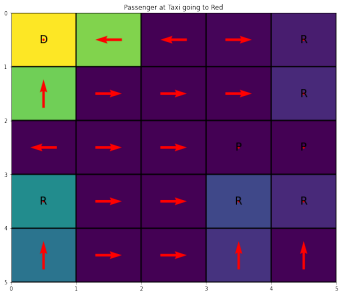


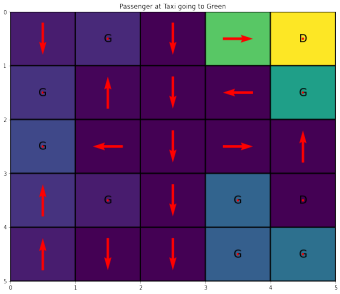


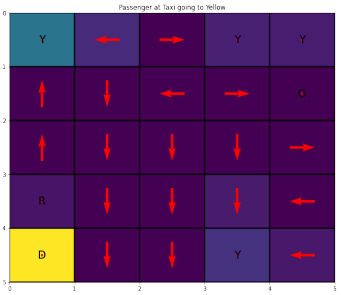


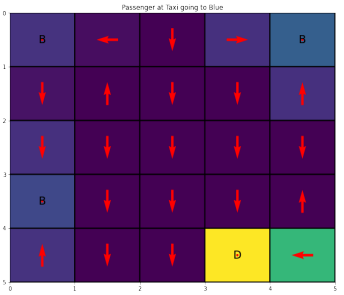












4 Description

Policies

Policy Description:

For SMDP Q-Learning the figures suggests that the learning agent takes the right next option, passenger position and destination position are inferred from the state. The sequence goes like first going to the passengers location, secondly picking up the passenger and then drop him in the destination location in least amount of time.
For Intra Option Q-Learning, the policies are defined such that each option could start at any cell in the grid and take the taxi to a particular location, the option terminates by choosing to pick up or drop off.

Alternate Options:

No Mutually exclusive set of options can improve the solution. As the taxi is bound to go to the designated four places to pick and drop a customer, the given option covers all the cases.

However improving on the given options can give better results and faster convergence. Some of the possible improvements are:

- Option to go from one designated place to other, 6 combinations.
- Initiating only the options to pick up location and drop location.

5 Inference

- Intra Option method converges faster and gives a better average reward, this is due to higher update frequency and efficient use of the State, Action, Reward, State samples.
- SMDP tend to have more option based action in Q-Table, whereas Intra Option tend to have primitive actions. This is because primitive Q action is also updated alongside options in intra option method.