

Report

CS6700 Spring 2024

Assignment 3

Prof Balaram Ravindran

Submitted by:

Akranth Reddy me20b100@smail.iitm.ac.in

Hemanth me20b045@smail.iitm.ac.in

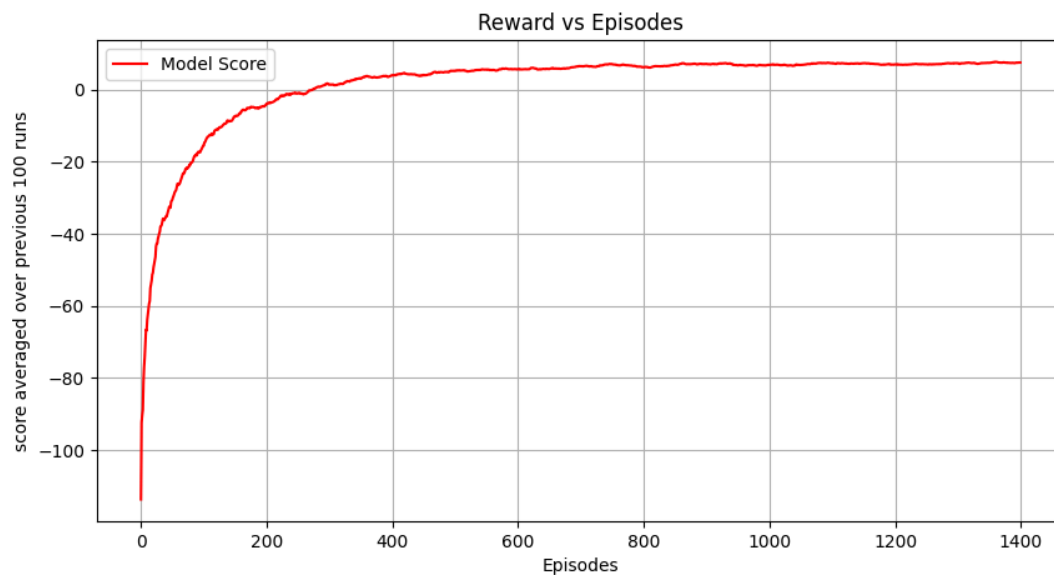
Date: 20-04-2024

[Github Link](#)

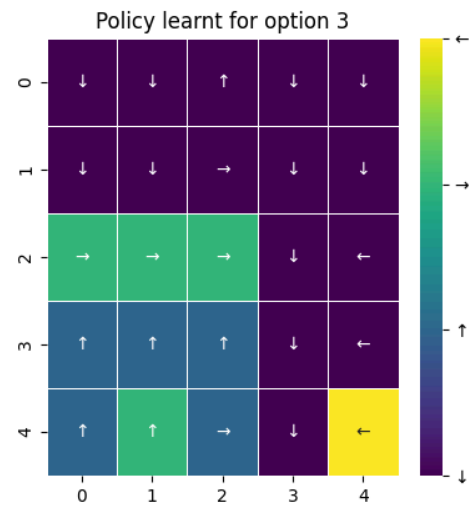
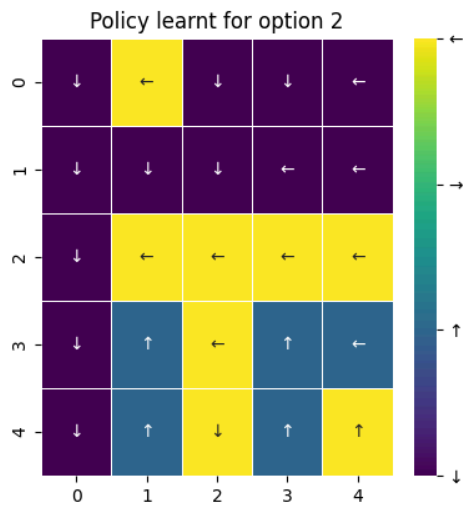
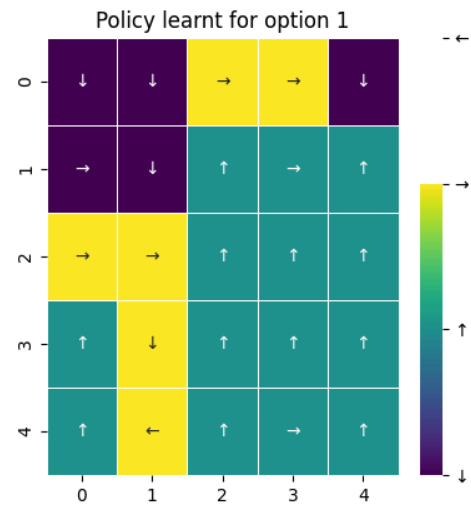
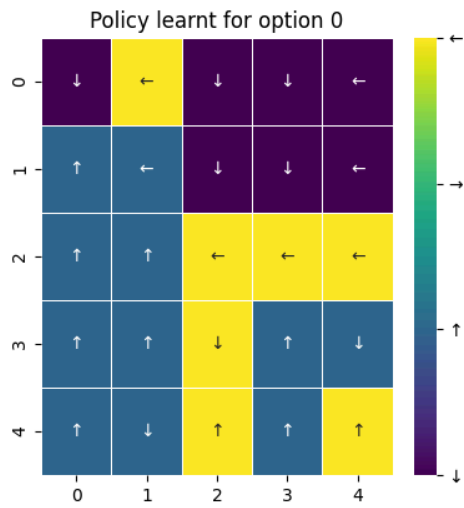
1.

SMDP Q-Learning

Reward Curves



Policy learnt for respective options

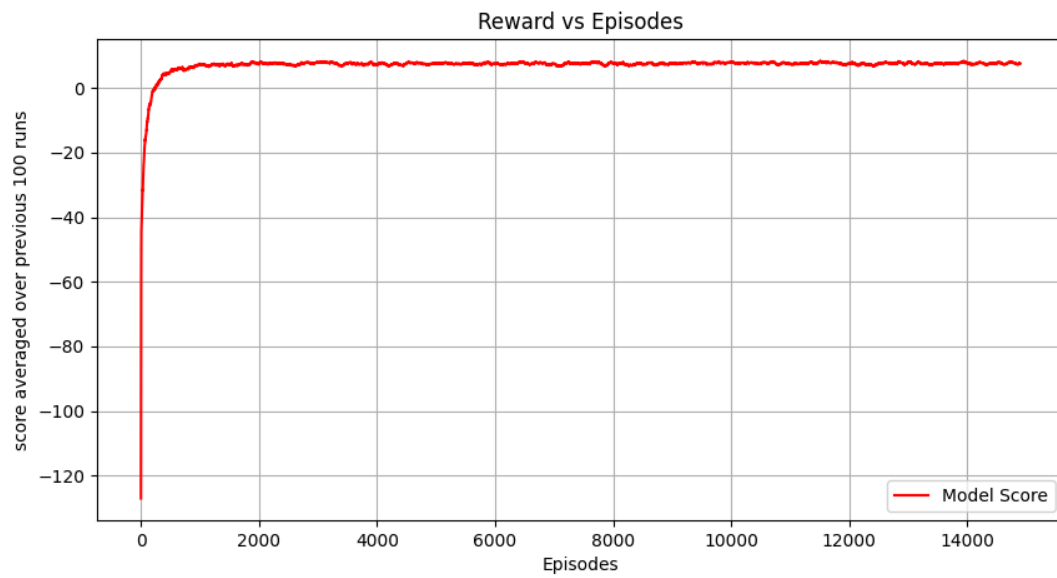


Q-Values:

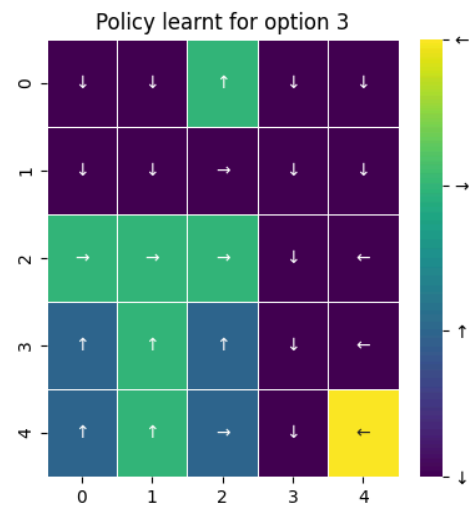
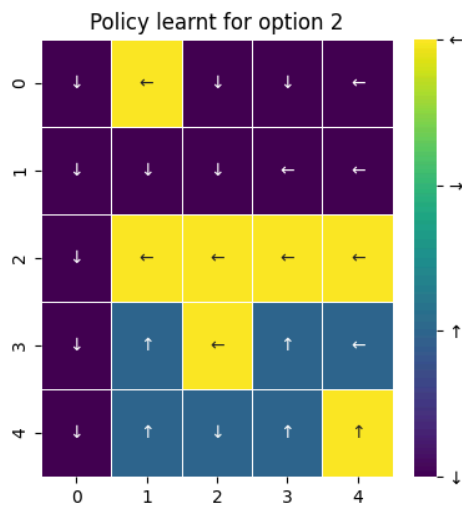
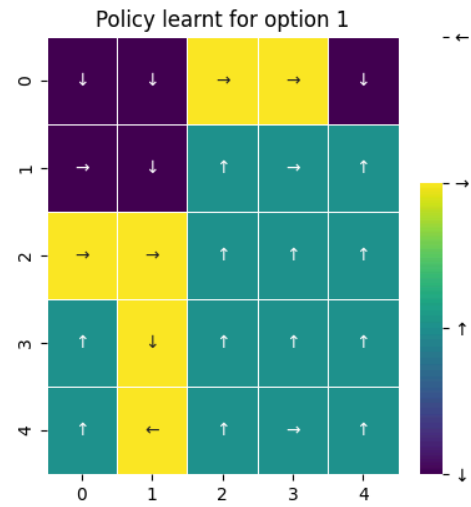
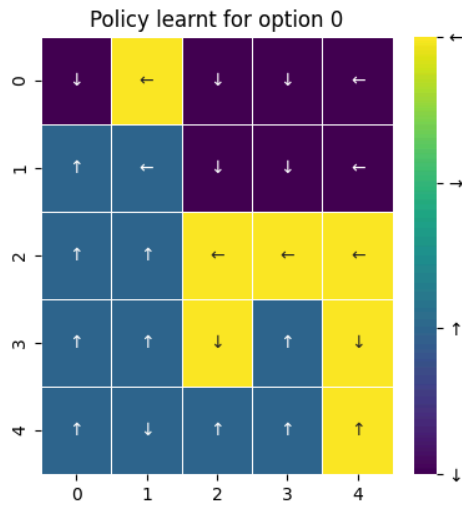


Intra-Option Q-Learning

Reward Curves



The policy learnt for respective option



Q-Values



2.

SMDP:

- The policy learnt for option 0 is to move east if the passenger is at location 1 or 2, and to move north if the passenger is at location 3. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move south to drop them off at location 0.

- The policy learnt for option 1 is to move west if the passenger is at location 0 or 3, and to move south if the passenger is at location 2. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move north to drop them off at location 1.
- The policy learnt for option 2 is to move north if the passenger is at location 0 or 1, and to move east if the passenger is at location 3. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move west to drop them off at location 2.
- The policy learnt for option 3 is to move south if the passenger is at location 0 or 2, and to move west if the passenger is at location 1. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move east to drop them off at location 3.

Intra-Option Q-Learning:

- The policy learnt for option 0 is to move east if the passenger is at location 1 or 2, and to move north if the passenger is at location 3. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move south to drop them off at location 0.
- The policy learnt for option 1 is to move west if the passenger is at location 0 or 3, and to move south if the passenger is at location 2. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move north to drop them off at location 1.
- The policy learnt for option 2 is to move north if the passenger is at location 0 or 1, and to move east if the passenger is at location 3. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move west to drop them off at location 2.
- The policy learnt for option 3 is to move south if the passenger is at location 0 or 2, and to move west if the passenger is at location 1. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move east to drop them off at location 3.

Reasoning:

- Both the SMDP and intra-option Q-learning algorithms learn the optimal policy for the Taxi environment.
- This is because both algorithms take into account the reward structure of the environment and learn to maximize the expected return.

- The main difference between the two algorithms is that the intra-option Q-learning algorithm can learn about the values of executing certain options without ever executing those options.
- This is because the intra-option Q-learning algorithm uses a special temporal-difference method to learn about the consequences of one policy while actually behaving according to another, potentially different policy.
- This allows the intra-option Q-learning algorithm to learn more efficiently than the SMDP algorithm.
- The SMDP algorithm learns the policy for each option by iteratively updating the Q-values for each state-action pair. The Q-values represent the expected return for taking a particular action in a given state. The algorithm then uses the Q-values to select the action with the highest expected return in each state.
- The Intra-Option Q-Learning algorithm also learns the policy for each option by iteratively updating the Q-values for each state-action pair. However, the Intra-Option Q-Learning algorithm also uses the Q-values to learn about the value of executing certain options without ever executing those options. This is because the Intra-Option Q-Learning algorithm can learn from the experience of other options.
- Both the SMDP and Intra-Option Q-Learning algorithms learn the optimal policy for the task of picking up and dropping off passengers in the Taxi environment. This is because both algorithms take into account the reward structure of the environment and learn to maximize the expected return.

Comparison between SMDP Q-Learning and intra-option Q-Learning algorithms:

SMDP Q-Learning:

- Learns the policy for each option by iteratively updating the Q-values for each state-action pair.
- Uses the Q-values to select the action with the highest expected return in each state.

Intra-option Q-Learning:

- Also learns the policy for each option by iteratively updating the Q-values for each state-action pair.
- Uses the Q-values to learn about the value of executing certain options without ever executing those options.
- Can learn from the experience of other options.

Improvement with intra-option Q-Learning:

- Yes, intra-option Q-Learning can improve performance over SMDP Q-Learning.

- This is because intra-option Q-Learning can learn from the experience of other options, which can help it to find a better policy.
- Intra-option Q-Learning can also learn about the value of executing certain options without ever executing those options, which can help it to make better decisions about which option to execute.

Briefly:

- SMDP Q-Learning learns the policy for each option independently.
- Intra-option Q-Learning can learn from the experience of other options.
- Intra-option Q-Learning can improve performance over SMDP Q-Learning.

3.

Alternative options