CS6700 - Reinforcement Learning
Programming Assignment 3

Argha Boksi(CS21D407),
Jashaswimalya Acharjee(CS22E005)

April 25, 2023

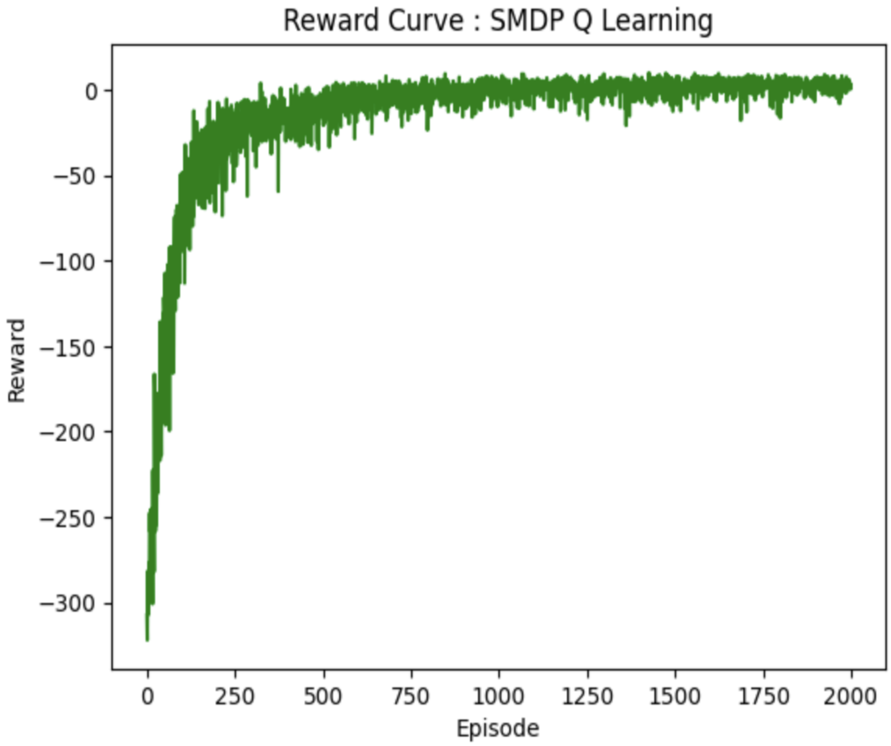
1 Options

- There are five possible passenger locations - 0: Red; 1: Green; 2: Yellow; 3: Blue; 4: in taxi
- And there are four possible destinations - 0: Red; 1: Green; 2: Yellow; 3: Blue
- Options are defined as to move the taxi from a given cell to each of the four possible destinations.
- Hence there will be four different options, one for each destination - **Go to red, Go to green, Go to yellow, Go to blue.**
- Option initiation states are all the states except when the taxi is at the designated location.
- Option policy is a deterministic optimal policy to reach the designated location.
- Option terminates once the taxi reaches the designated location.
- There are 6 discrete deterministic actions**(0: move south; 1: move north; 2: move east; 3: move west; 4: pick passenger up; and 5: drop passenger off)** and 4 options**(Go to red, Go to green, Go to yellow, Go to blue)**, therefore we have a maximum of 10 possible options to execute at any state.

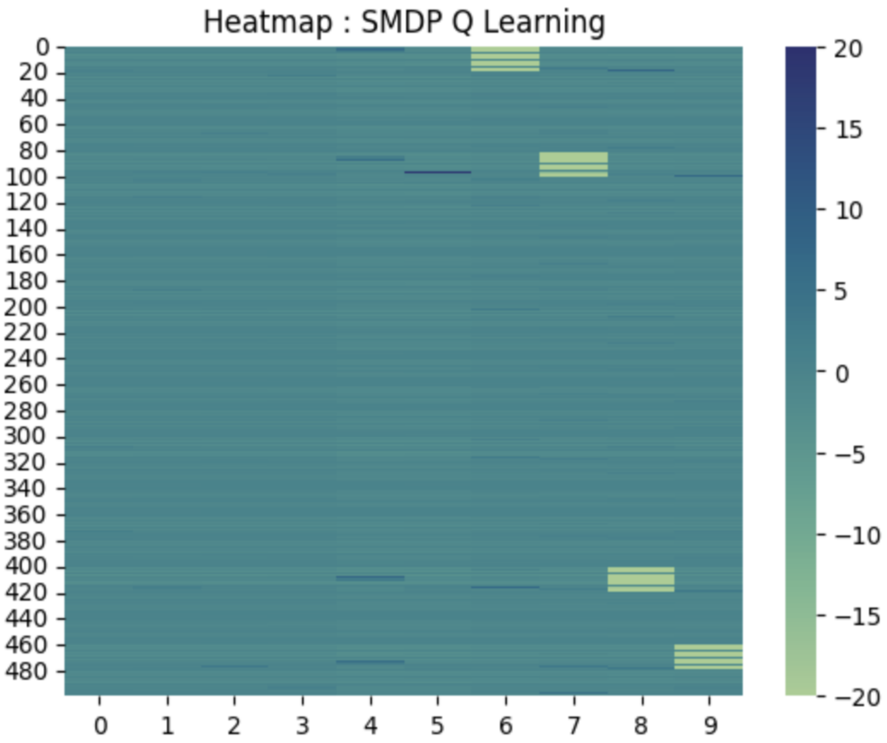
2 SMDP Q-Learning

- We implemented single step SMDP Q-learning for solving the taxi problem.
- We found that the following hyperparameter setting gave best results - $\gamma = 0.9$, $\epsilon - start = 1.0$, $\epsilon - decay = 0.9$, $\alpha = 0.1$
- For the above hyperparameter setting we ran 10 experiments and averaged the results to generate reward curve and heatmap of the learned Q-values.

2.1 Reward Curve



2.2 Visualization of learned Q-values



2.3 Policies Learnt

- The agent first executes the option to move the taxi to the pickup location.
- Then the agent executes pickup primitive action.
- Then the agent executes the option to move the taxi to the destination.
- Then the agent executes the dropoff primitive action

2.4 Some Observations

- There are many state-option pairs that have not received enough updates. So for those states agent often ends up taking random actions.
- In SMDP Q-learning we update the Q-table once the option execution has terminated. During option execution the Q-values for the visited state-option pairs are not updated.
- Hence we find that many intermediate state-option pairs do not receive any Q-value updates.
- However, many of the state-option pairs received frequent updates. In those states the agent has learned how to behave optimally.

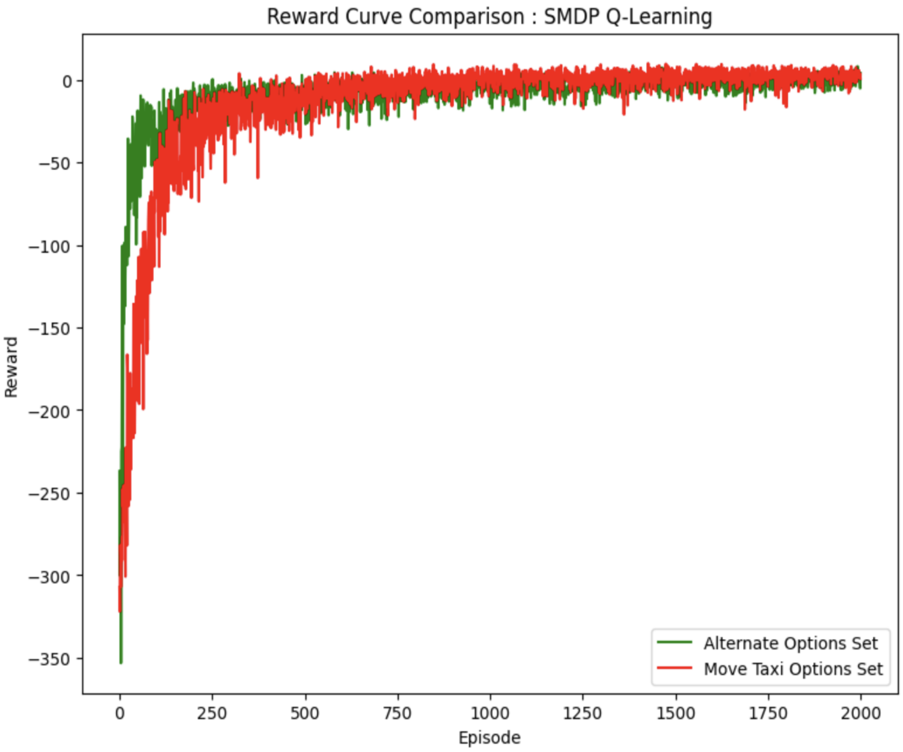
2.5 Alternate Set of Options

- We have defined eight alternate options.
- Corresponding to every designated location we have two options.
- For designated location "Green" the two options are - (a) Go to location Green(G) and pick up, (b) Go to location Green(G) and drop off.
- Similarly we have two options defined for each of the three other locations(Red, Yellow, Blue).
- Option initiation states: all the states belong to initiation set.
- Option policy: optimal policy is used to move a taxi to the designated location and then either pickup or drop-off is executed.
- Option termination condition: Once either pickup or drop-off action is executed the option terminates. For example, option "Go to location green and drop-off" terminates when taxi goes to location green and executes drop-off action.

2.6 SMDP Q-learning with Alternate Option Set and Move taxi Option set

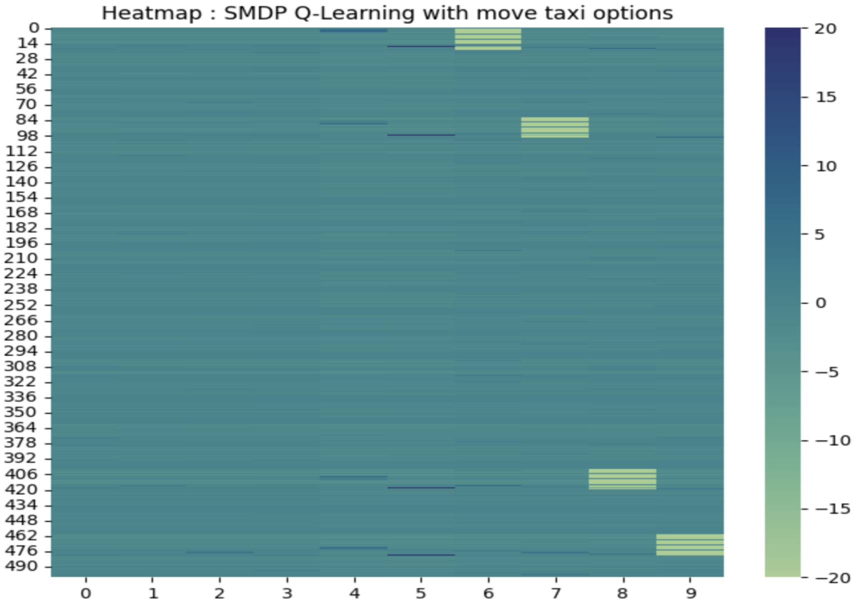
- We ran SMDP Q-learning with both alternate option set and the given move taxi option set.
- We present below the reward curve comparison and the Q-value heatmap comparison.

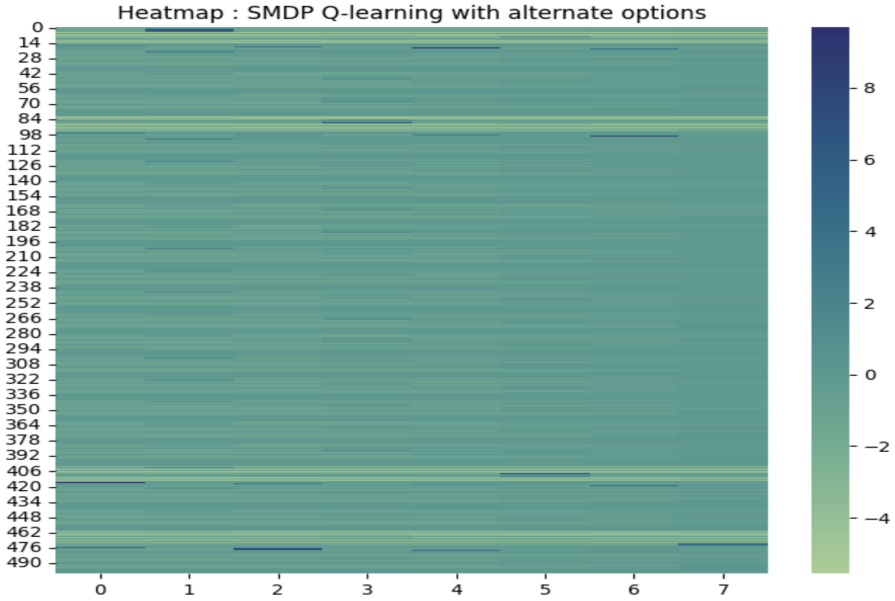
2.6.1 Reward Curve Comparison



From the above figure it can be seen that SMDP Q-learning with alternate options set converges faster than SMDP Q-learning with the given move taxi options set.

2.6.2 Q-value heatmap comparison

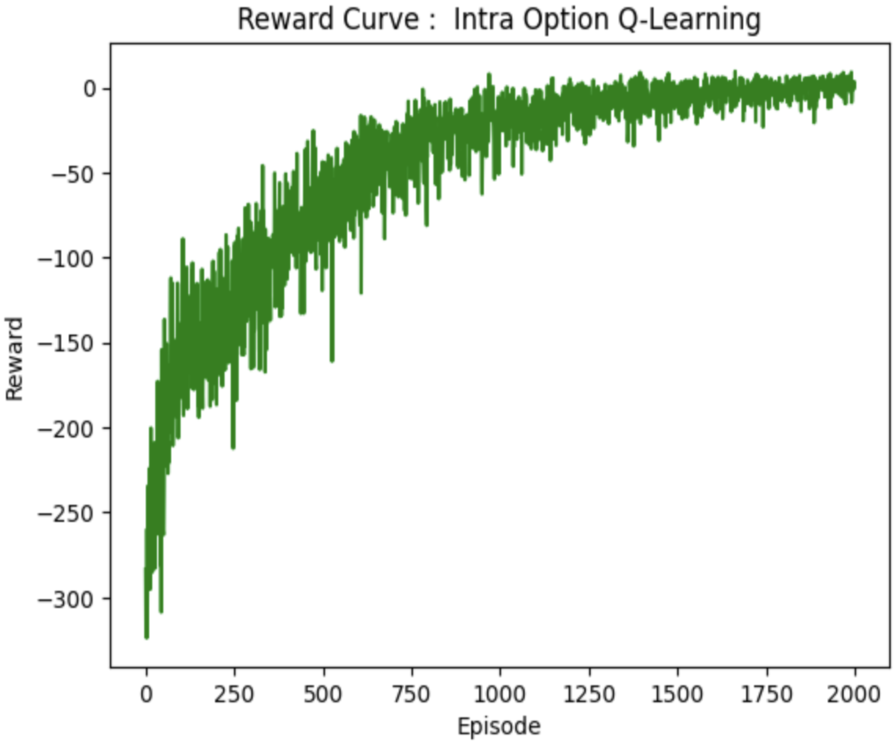




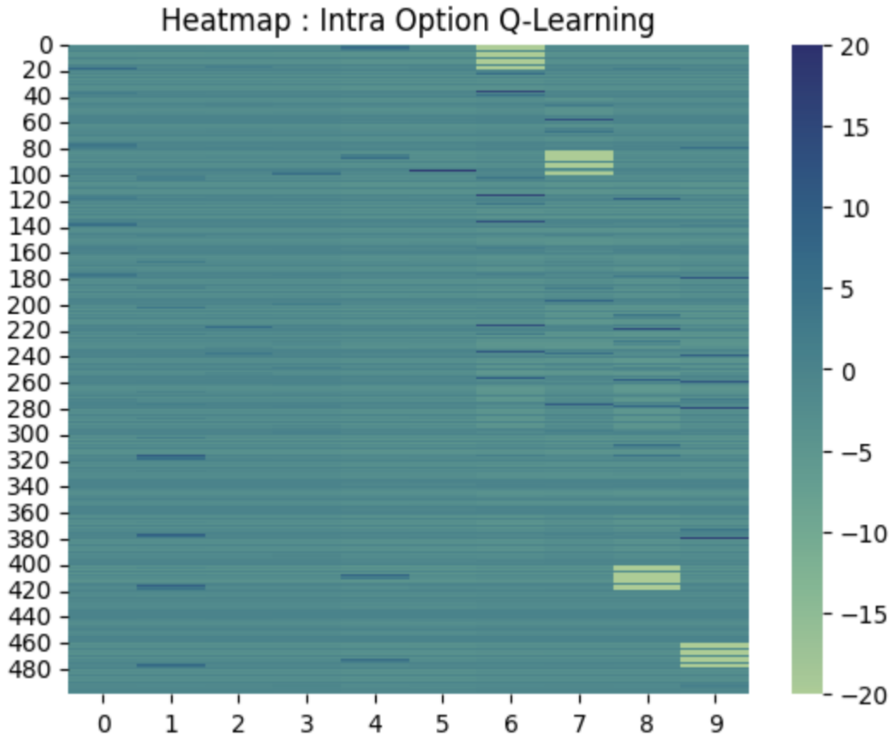
3 Intra-Option Q-Learning

- We implemented Intra-option Q-learning for solving the taxi problem.
- We found that the following hyperparameter setting gave best results - $\gamma = 0.9$, $\epsilon - start = 1.0$, $\epsilon - decay = 0.9$, $\alpha = 0.1$
- For the above hyperparameter setting we ran 10 experiments and averaged the results to generate reward curve and heatmap of the learned Q-values.

3.1 Reward Curve



3.2 Visualization of learned Q-values



3.3 Policies Learnt

- The agent first executes the option to move the taxi to the pickup location.
- Then the agent executes pickup primitive action.
- Then the agent executes the option to move the taxi to the destination.
- Then the agent executes the dropoff primitive action

3.4 Some Observations

- The number of updates for intra option Q-learning is greater than SMDP Q-learning.
- Hence most of the state-option pairs in the Q-table have received frequent updates.
- This means we are more likely to see the optimal policy being executed at different states.

3.5 Alternate Set of Options

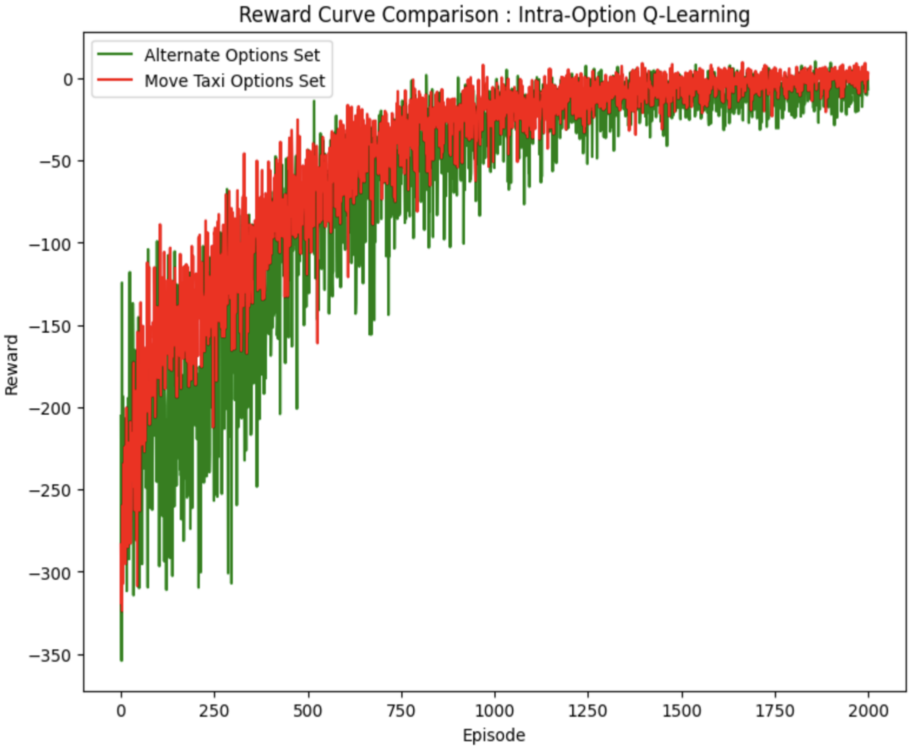
- We have defined eight alternate options.
- Corresponding to every designated location we have two options.

- For designated location ”Green” the two options are - (a) Go to location Green(G) and pick up, (b) Go to location Green(G) and drop off.
- Similarly we have two options defined for each of the three other locations(Red, Yellow, Blue).
- Option initiation states: all the states belong to initiation set.
- Option policy: optimal policy is used to move a taxi to the designated location and then either pickup or drop-off is executed.
- Option termination condition: Once either pickup or drop-off action is executed the option terminates. For example, option ”Go to location green and drop-off” terminates when taxi goes to location green and executes drop-off action.

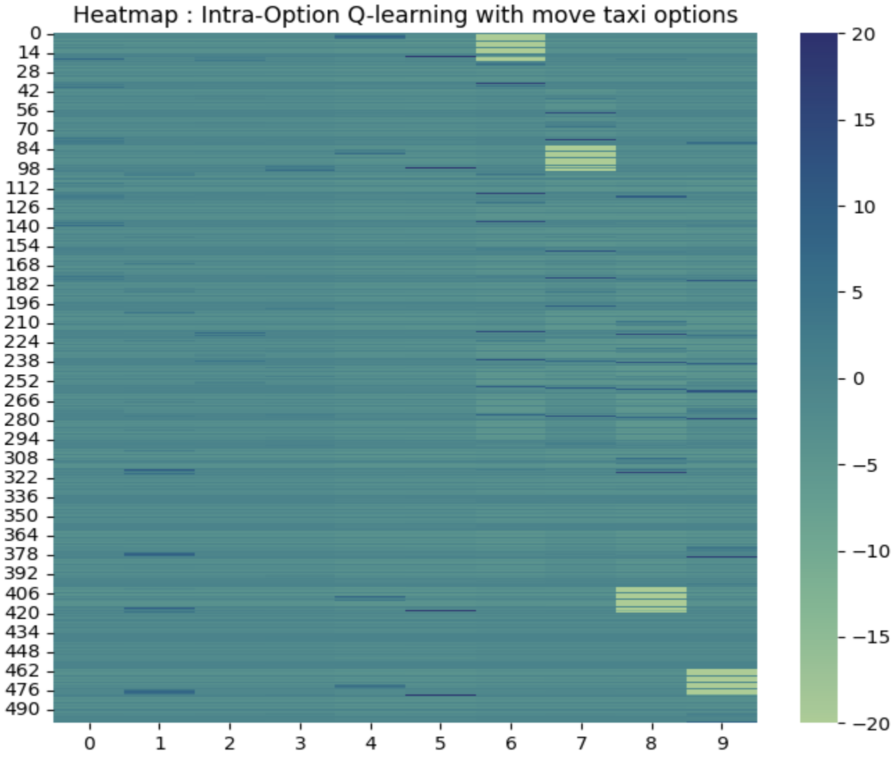
3.6 Intra-Option Q-learning with Alternate Option Set and Move taxi Option set

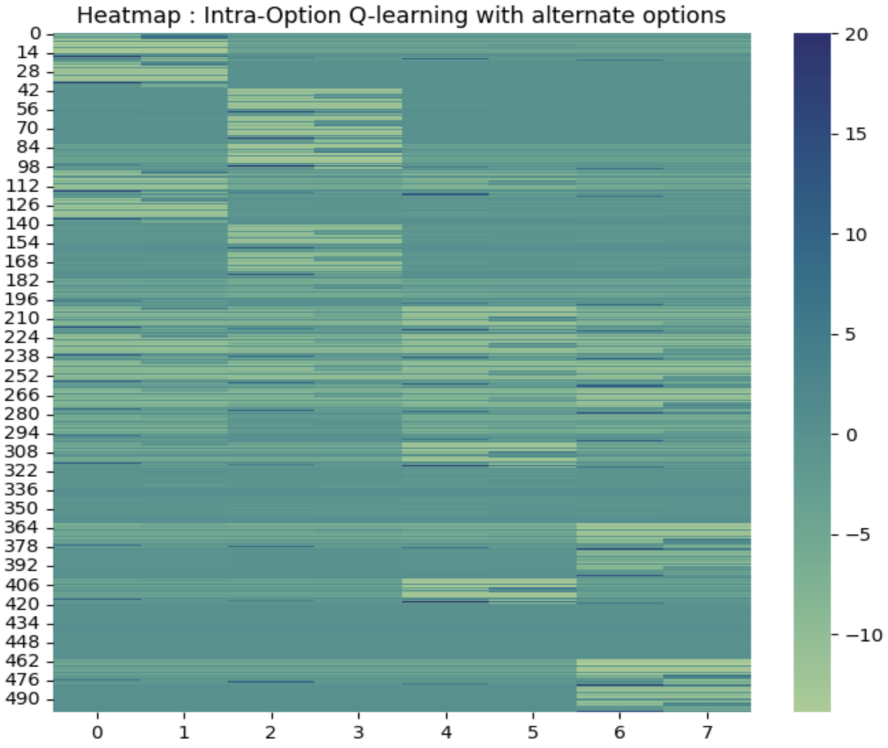
- We ran Intra-Option Q-learning with both alternate option set and the given move taxi option set.
- We present below the reward curve comparison and the Q-value heatmap comparison.

3.6.1 Reward Curve Comparison



3.6.2 Q-value Heatmap Comparison





4 Comparison between SMDP Q-Learning and Intra-Option Q-Learning

- **Observations:** Here we describe our learning from the experiments conducted for Option type 1:
  - For option type 1, both the SMDP Q Learning agent and the Intra-option Q Learning agent, learn to reach an R/G/B/Y cell from another R/G/B/Y cell using the shortest possible time-steps, i.e., in an optimal manner. As shown in the plots, both the agents are able to figure out the utility of picking these options defined by us over choosing primitive actions defined by the environment.
  - As per Q table visualisations for both the algorithms, intra-option Q learning makes more number of updates to various states encountered during training, as compared to SMDP Q learning. This behaviour is expected because in the SMDP model-learning method, Q-value updates are done after the termination of an option with the accumulated reward, whenever it is selected, whereas, in the intra-option model-learning method, Q-value updates are done at every step to all options that are consistent with the action taken on that step.
  - From the reward comparison plots for both the type of options, we can conclude that intra-option Q Learning method performs *slightly* better (converges to higher reward) as compared to SMDP Q Learning method.

  Next, we describe our learning from the experiments conducted for Alternate Options and compare the results below:

  - For Alternate options, both the SMDP Q-Learning agent and the Intra-Option Q-Learning agent, learn to move in a particular direction from the current state and ultimately reach an R/G/B/Y cell. The path discovered may not be optimal, as suggested by the cell-visit count plot for this set of options, but the agent explores the grid more. We can also see that the agent visits the middle row cells quite often to reach other cells. As shown in the heatmaps, it also visits the cells lying between two walls initially while exploring the environment, as compared to the first option type. Both the agents are eventually able to figure out the utility of picking these options defined by us over choosing primitive actions defined by the environment.
  - We observe from the comparitive plots that intra-option Q learning converges slightly faster using the initial options as compared to when learning with Alternate options, which makes sense because the first type is defined to follow shortest paths between cells.