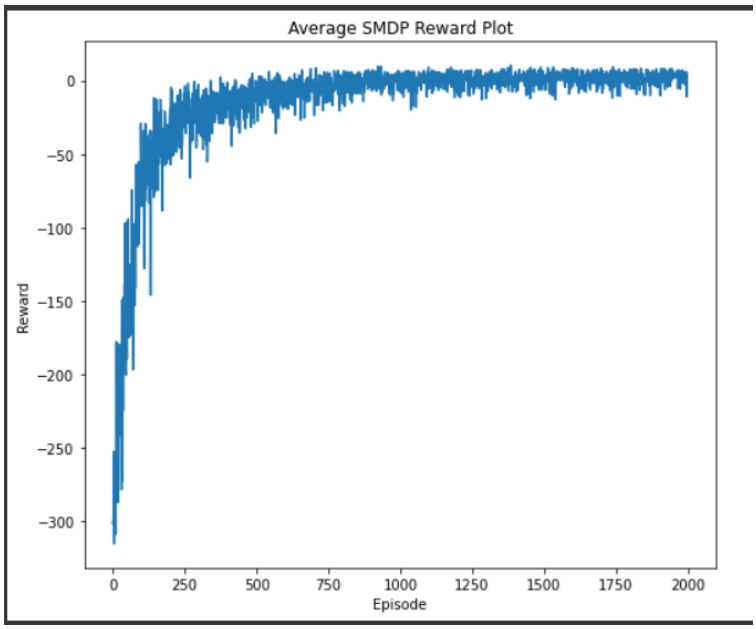
CS6700: Reinforcement Learning
Assignment #3

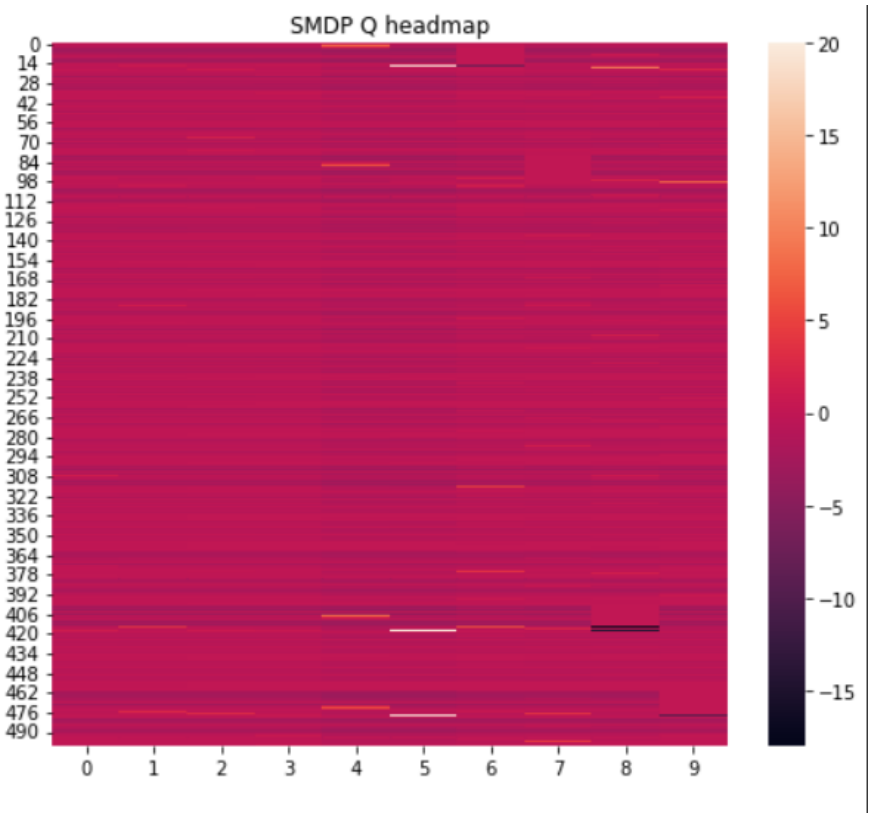
Topics: Hierarchical Reinforcement Learning **Deadline:** 21 April 2022

Teammate 1: (AKASH SAINI) **Roll number:** CS21M003
Teammate 2: (RETURAJ BURNWAL) **Roll number:** CS21D406

- Options:
  - There are 4 designated locations RED, BLUE, GREEN and YELLOW. Options are defined as to move the taxi from a given cell to each of the designated locations. Hence there will be 4 different options, 1 for each locations.
  - Options 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 termination condition: Once the taxi reaches the designated location, option terminates.
- There are 6 primitive actions and 4 options, therefore we have maximum of 10 possible options to execute at any state.
- SMDP Q-Learning
  - Solution 1:
    - Reward Curve:



* Heatmap of Q values:



– Solution 2:

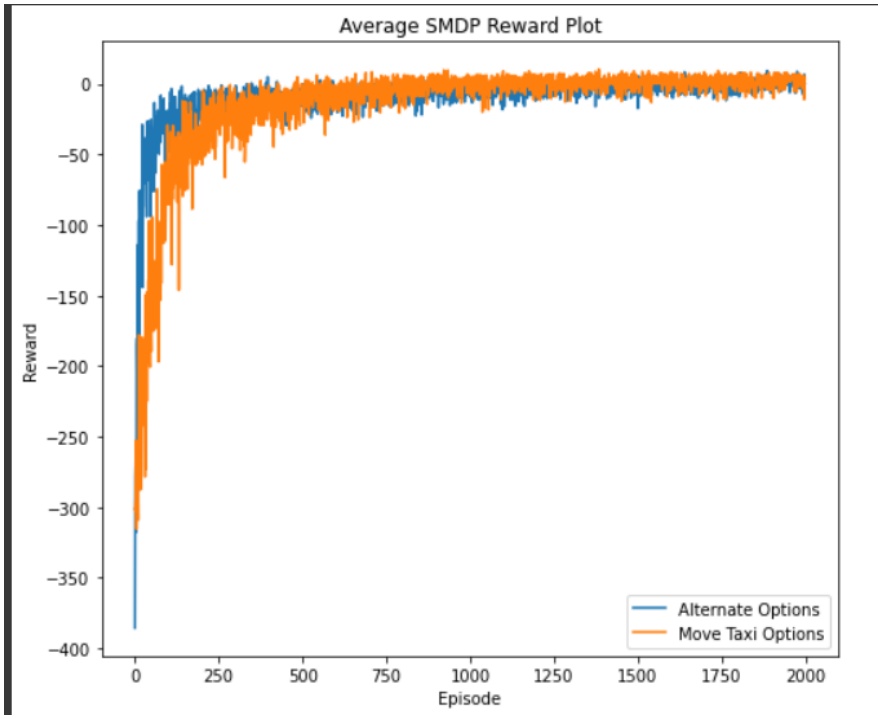
* Learned Policy: It first executes the option to move the taxi to the pickup location, then executes pickup action which is followed by executing the option to move the taxi to the dropoff location and then executes the dropoff action.
* There are still many states-option pairs that have not been updated. So for those states agent executes some random policy.
* Reason: SMDP updates the Q table only when option execution has ended. Since during option execution, the visited states are not updated this may lead to sparse values in the Q table. Hence for many state-option pairs we find that the Q table has not been updated.
* Those state-option pairs that have received an update have learned to behave optimally.

– Solution 3:

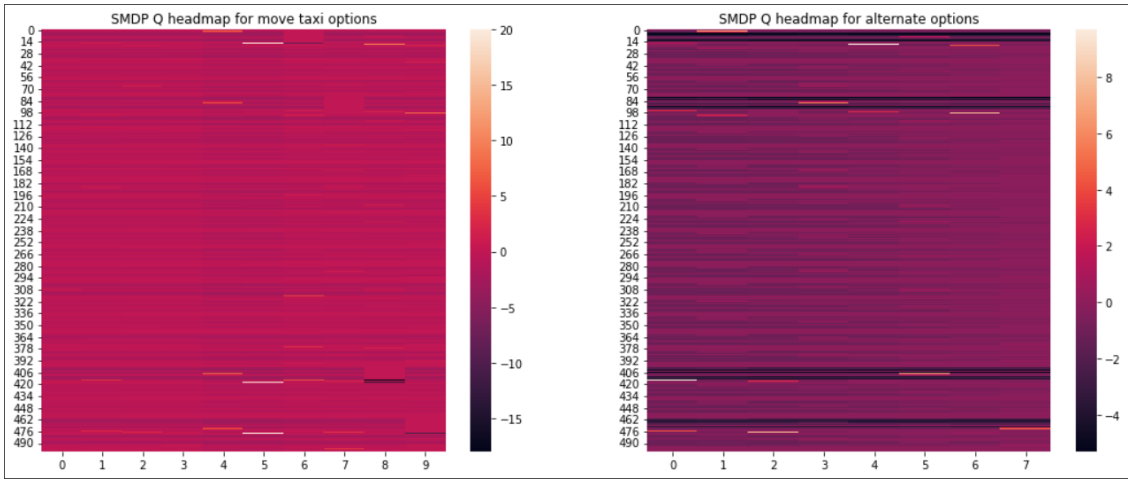
* Alternate Options:
  · We define alternate options set as: to move the taxi to each designated locations and then execute either pickup or drop off action.
  · Since there are 4 designated location and for each location pickup or dropoff action can be executed, so in total there are 8 options, for example alternate_option 1 can be "goto location red and drop-off".

- 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 "goto location red and drop-off" terminates when taxi goes to location red and executes drop-off action.

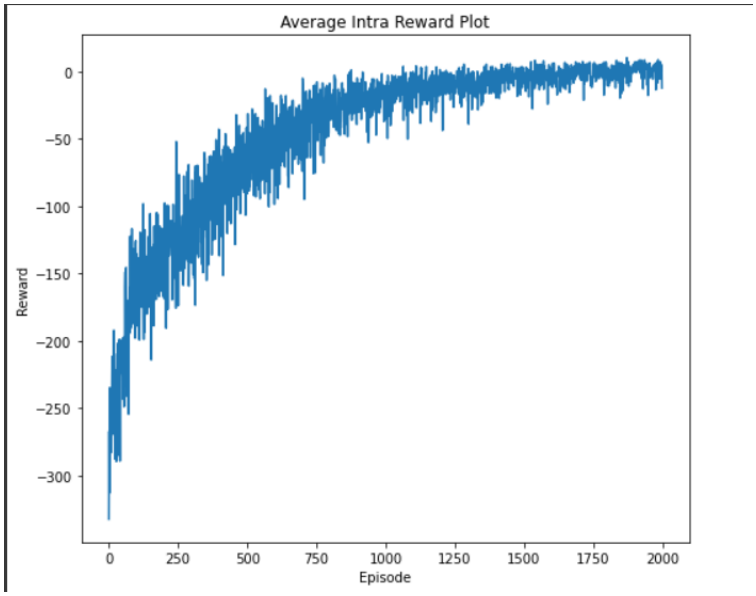
* We noticed that with only the alternate options, agent is able to complete the task. So for alternate option settings we have maximum of 8 possible options at any state.
* Comparing rewards: We noticed that alternate option set converges faster than the given option set.



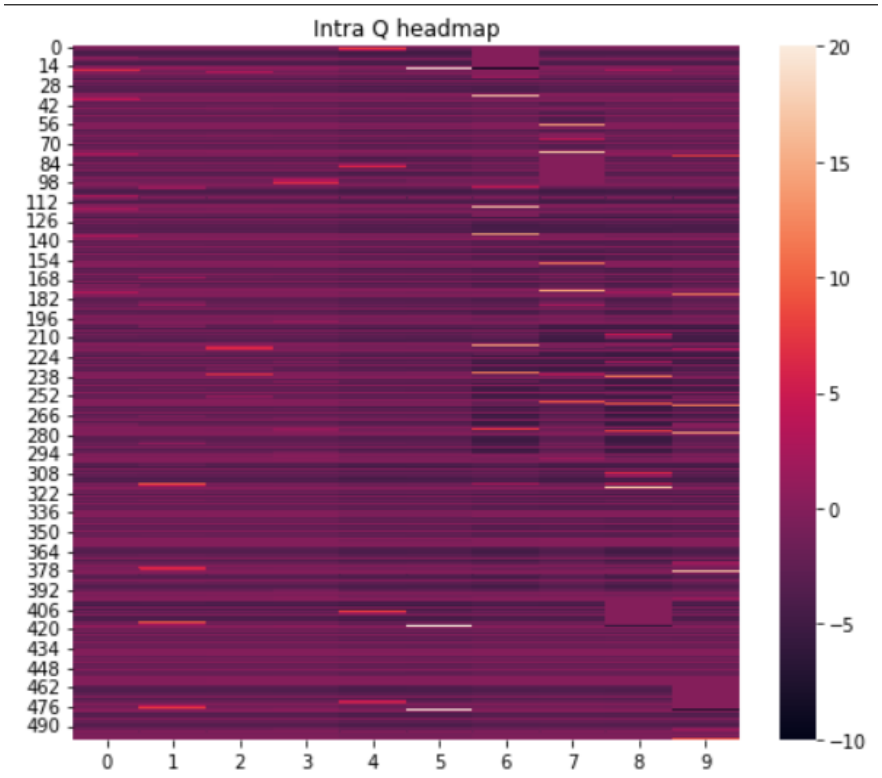
* Heatmap of Q value:



- Intra Option Q-Learning
  - Solution 1:
    - Reward Curve:



    - Heatmap of Q values:

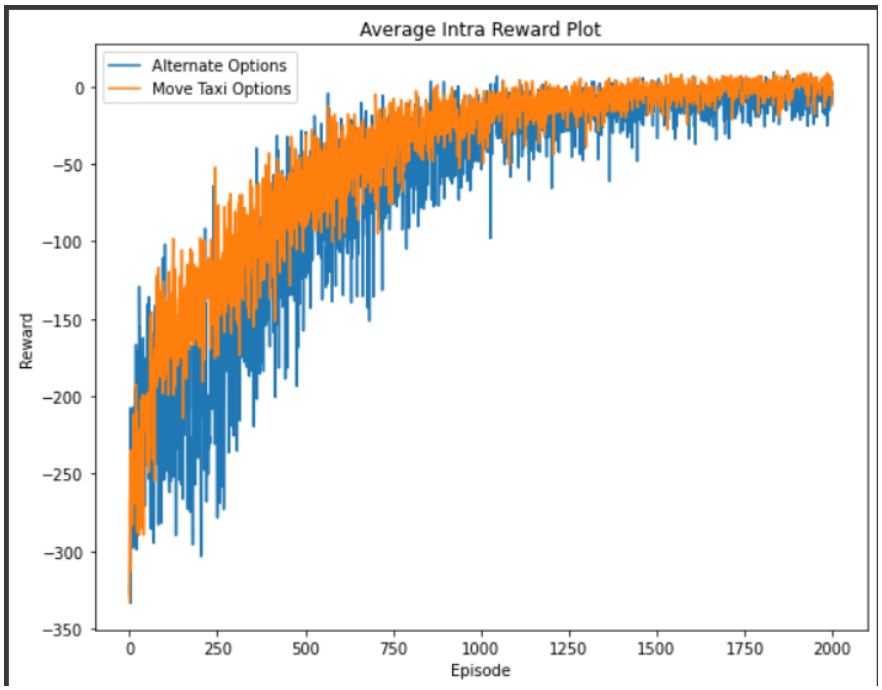


– Solution 2:

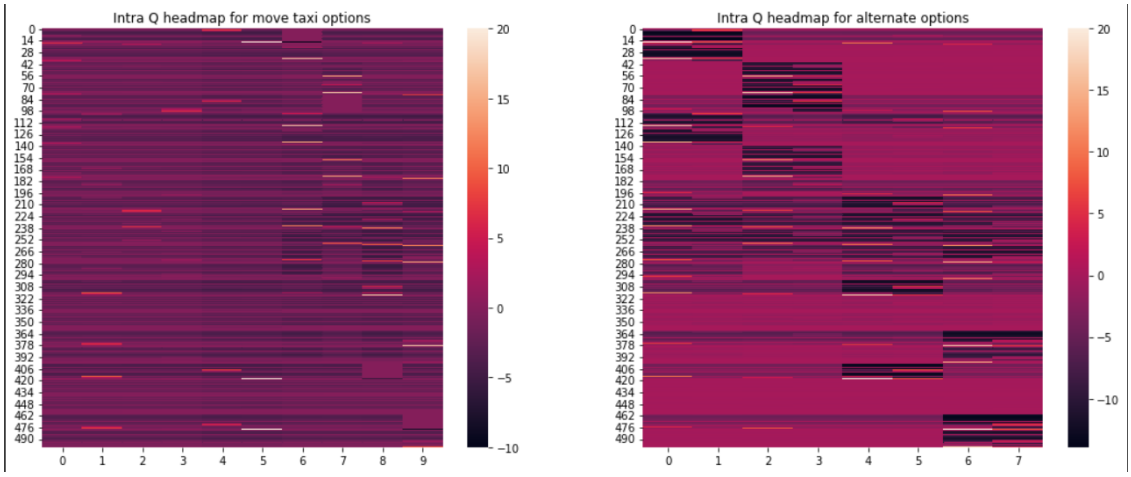
  * Learned Policy: It first executes the option to move the taxi to the pickup location, then executes pickup action which is followed by executing the option to move the taxi to the dropoff location and then executes the dropoff action.
  * Reason: Here most of the state-options pairs in the Q table have been updated and hence we are more likely to see the optimal policy being executed at different states.

– Solution 3:

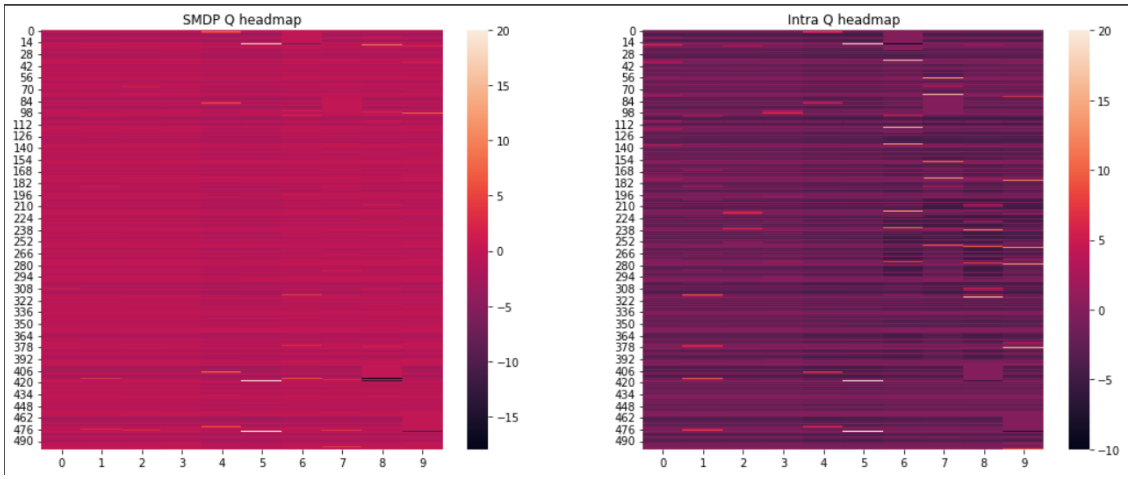
  * Alternate Options: Same as defined in the previous point under SMDP Alternate Options.
  * Comparing rewards: We noticed that alternate option set converges faster than the given option set.



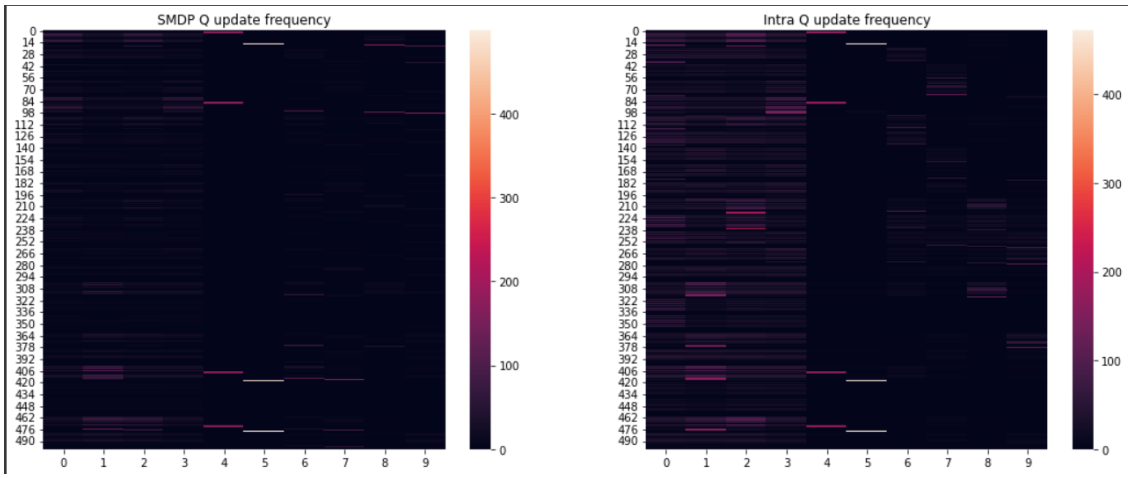
∗ Heatmap of Q value:



- Comparing Intra-Option and SMDP Q-Learning algorithm based on given option set i.e. to move the taxi to each designated locations.
  - Heatmap of Q values:



– Average number of state-option pair update plot:



– From the both the above plots we can notice that a significant number of state-option pairs in Intra-Option Q-Learning has received updates compared to SMDP Q-Learning. This is because SMDP updates its Q values only after option completes its execution where as Intra-Option Q-Learning updates intermediate state-option pairs during the execution of the option.

- Code link: Colab Notebook
