Reinforcement Learning Course Projects

Partha Sarathi Mohapatra
EE18D703

21 July 2020

Hierarchical Reinforcement Learning

In this Hierarchical Reinforcement Learning implementation there are in total 12 options (8 multi-step options and 4 primitive options). But if we take into consideration initiation states ($\mathcal{I}$) and the termination conditions ($\beta(s)$) of each multi-step options, then only 2 multi-step options (so in total 6 options including the primitive ones) will be valid in each state.

The size of $Q(s, o)$ table will be $|\mathcal{S}| \times |\mathcal{O}|$, where $\mathcal{S}$ and $\mathcal{O}$ represents the state space and the set of options respectively. So for 12 options the required memory size is $|\mathcal{S}| \times 12$. But as discussed above in each state only 6 options are valid, so only $|\mathcal{S}| \times 6$ size of memory is sufficient for storing all Q values (if we consider only the 2 valid multi-step options in each state as either clockwise or anticlockwise options, same indices can represent different multi-step options depending upon the initiation states $\mathcal{I}$).

1 SMDP Q-learning

We implement the SMDP Q-learning with eight predefined (multi-step) options for the two goal states G1 and G2 where the starting state was chosen randomly in the room 1. To visualize the learned Q- values we plot the heat maps for four primitive actions and multi-step options (combining all clockwise multi-step options as one and all anticlockwise multi-step options as another) in the following figures.

Heat maps for goal G1:

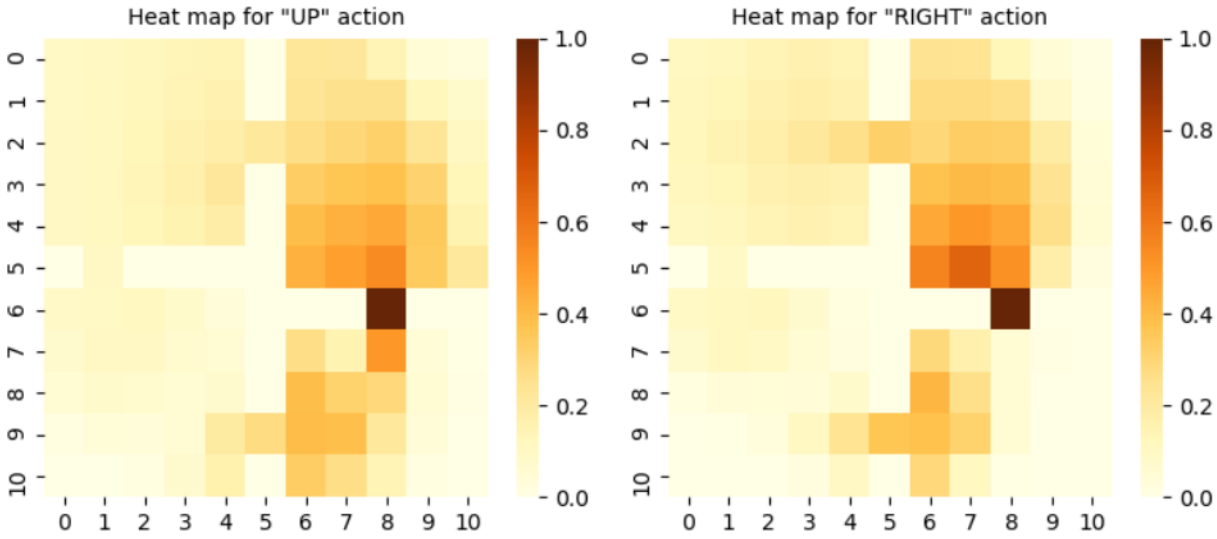

Figure 1: Heat map showing Q-values for "UP" and "RIGHT" actions for goal G1

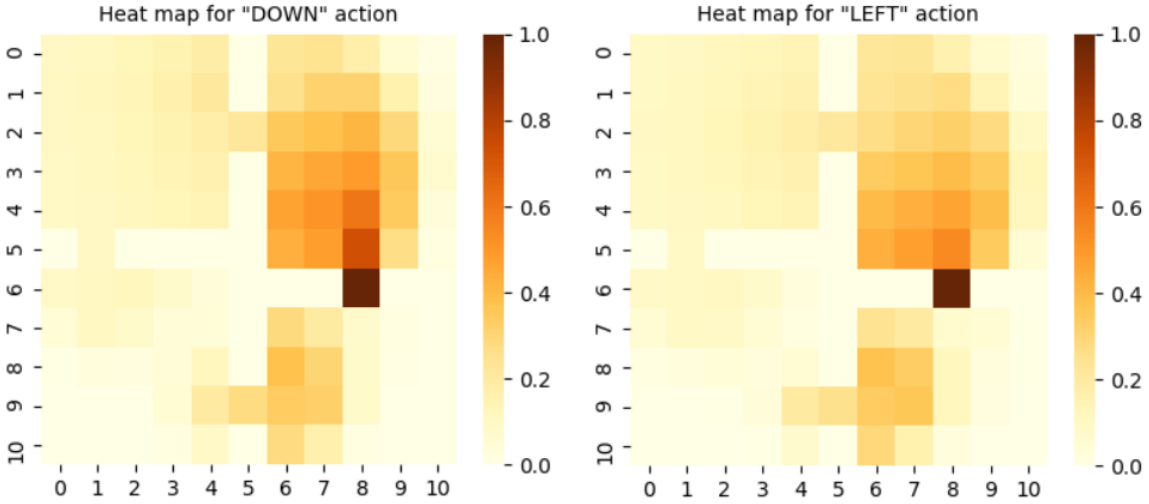

Figure 2: Heat map showing Q-values for "DOWN" and "LEFT" actions for goal G1

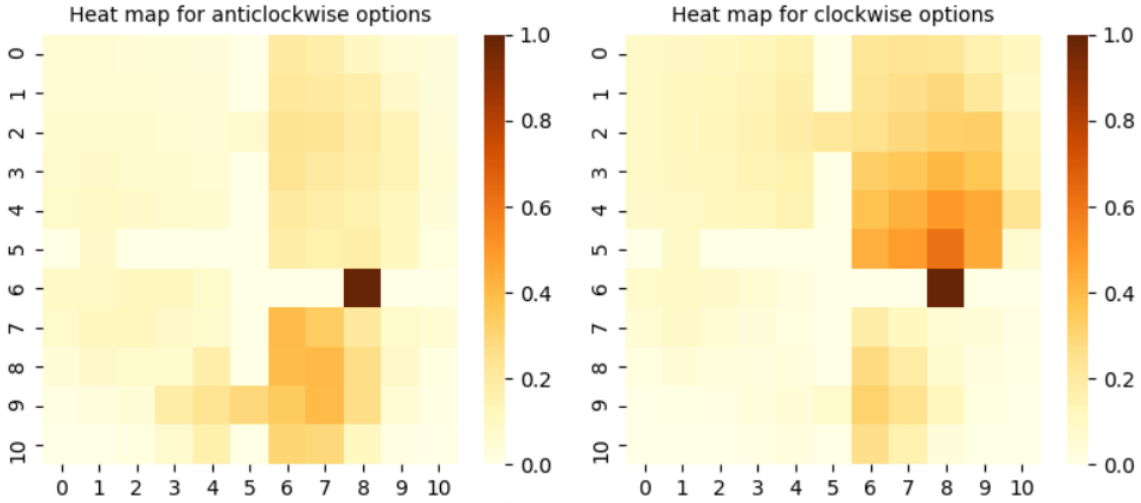

Figure 3: Heat map showing Q-values for anticlockwise and clockwise options for goal G1

Heat maps for goal G2:

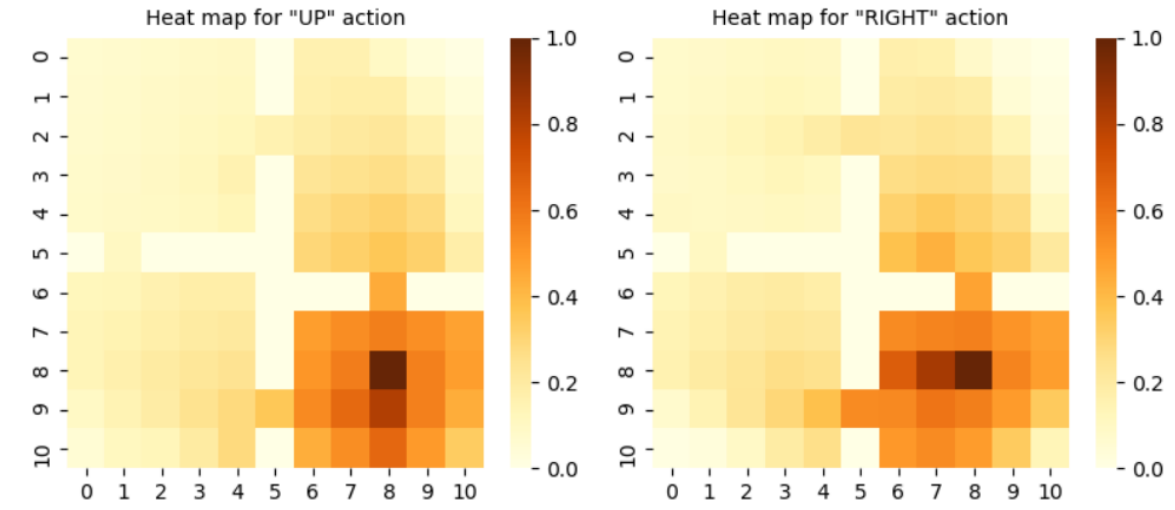

Figure 4: Heat map showing Q-values for "UP" and "RIGHT" actions for goal G2

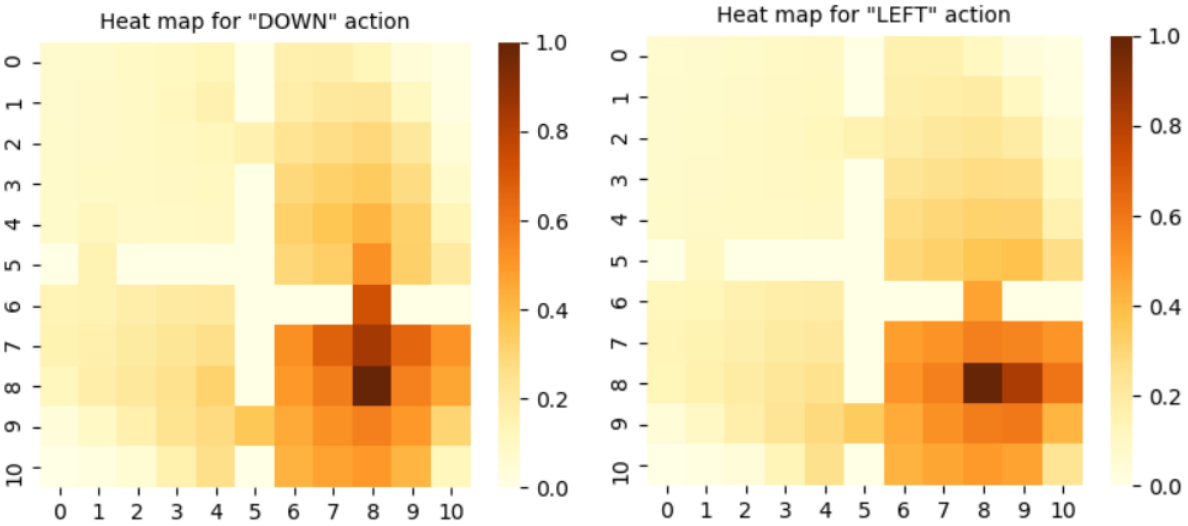

Figure 5: Heat map showing Q-values for "DOWN" and "LEFT" actions for goal G2

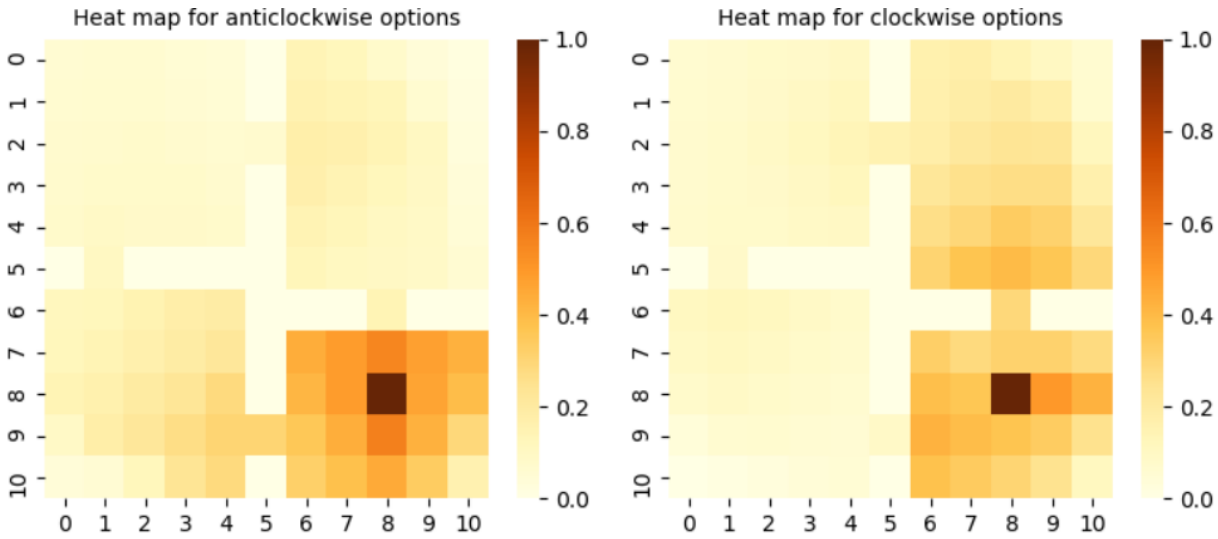

Figure 6: Heat map showing Q-values for anticlockwise and clockwise options for goal G2

1.1 Observations

(i) As can be seen from the above heat maps, the goal state has the highest Q-value (value of +1) ; as the only reward is +1, which the agent can get after reaching the goal.

(ii) Other states Q-values not only depend upon the number of steps needed to reach the goal but also the option taken in that state. For example in figure 2 and 5 for "DOWN" action the states which will directly lead to the goal state with "DOWN" action (the states directly above the goal) has higher Q-values than other nearer state for which the same action will not directly lead to the goal. For other options also similar patterns can be observed.

(iii) **Initial state in center of room** 4:(Heat maps for this case are shown in figures 7-12) If the initial state of the agent is changed to the center of room 4 (state 90), we see that for goal G1, most of the times, the agent first takes the anticlockwise multi-step option from room 4 to hallway of room 3 (state 77) and then the multi-step option to hallway of room 2 (state 56), which is also G1. For goal G2, the agent takes the same multi-step options to state 77 and then to state 56, thereafter it takes primitive actions to reach the goal G2 (state 64). These are represented abstractly below:
For goal G1:
$90 \xrightarrow{\text{multi-step option}} 77 \xrightarrow{\text{multi-step option}} 56$
For goal G2:
$90 \xrightarrow{\text{multi-step option}} 77 \xrightarrow{\text{multi-step option}} 56 \xrightarrow{\text{primitive actions}} 64$

We repeat the above implementation for both goal states G1 and G2 by changing the initial state to the center of the room 4 and plotted the learned Q- values in the figures shown below:

Heat maps for goal G1 with initial state in the centre of room 4:

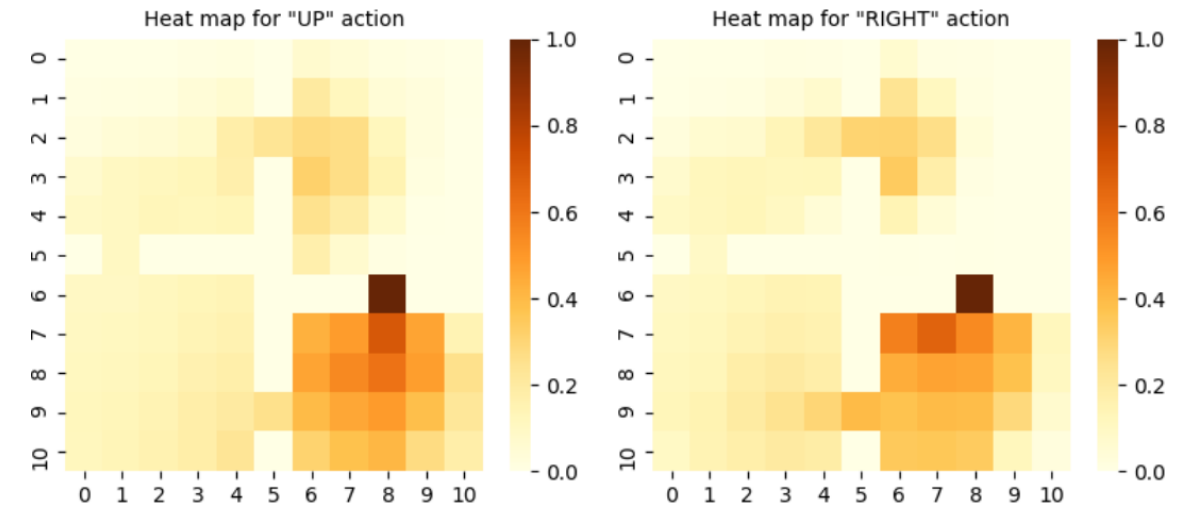

Figure 7: Heat map showing Q-values for "UP" and "RIGHT" actions for goal G1 with initial state in center of room 4

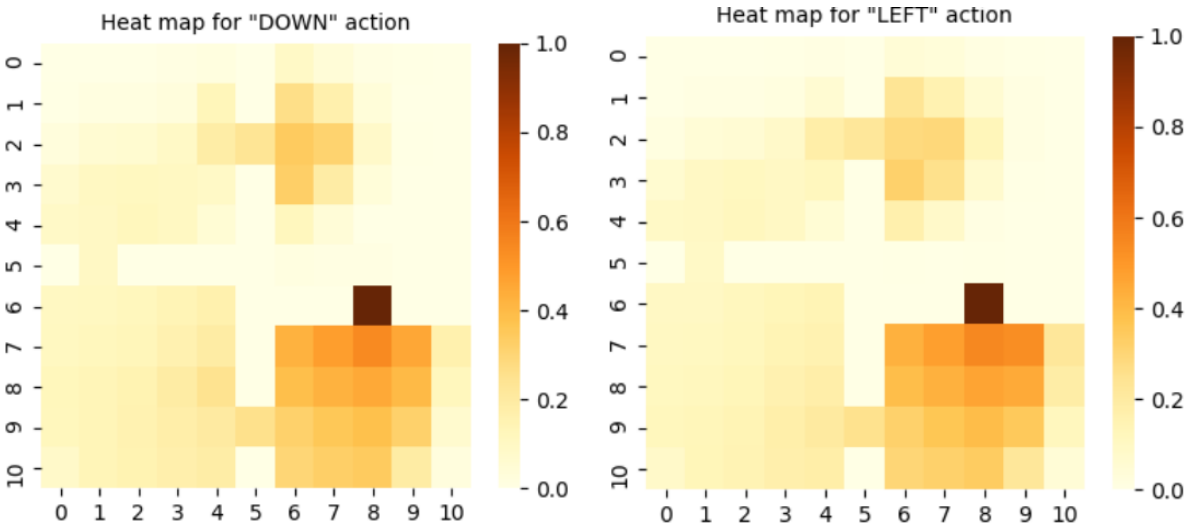

Figure 8: Heat map showing Q-values for "DOWN" and "LEFT" actions for goal G1 with initial state in center of room 4

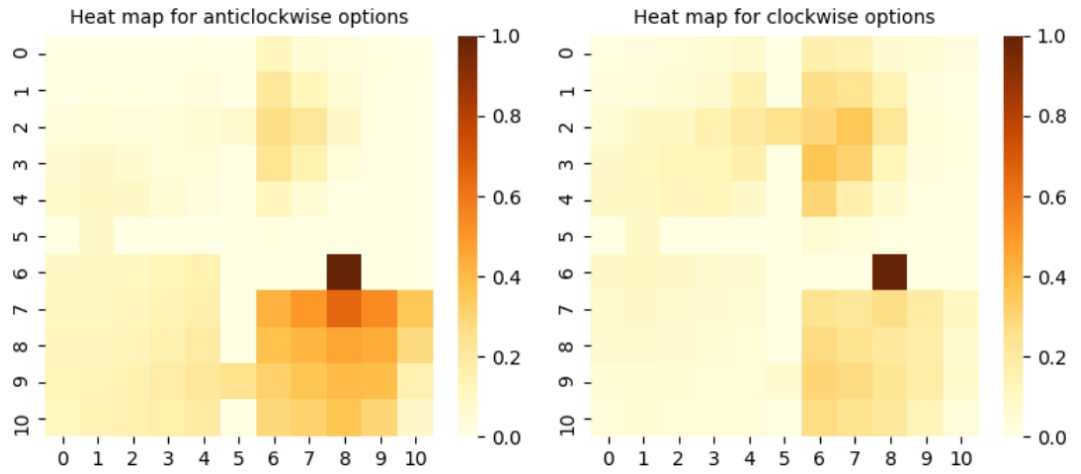

Figure 9: Heat map showing Q-values for anticlockwise and clockwise options for goal G1 with initial state in center of room 4

Heat maps for goal G2 with initial state in the centre of room 4:

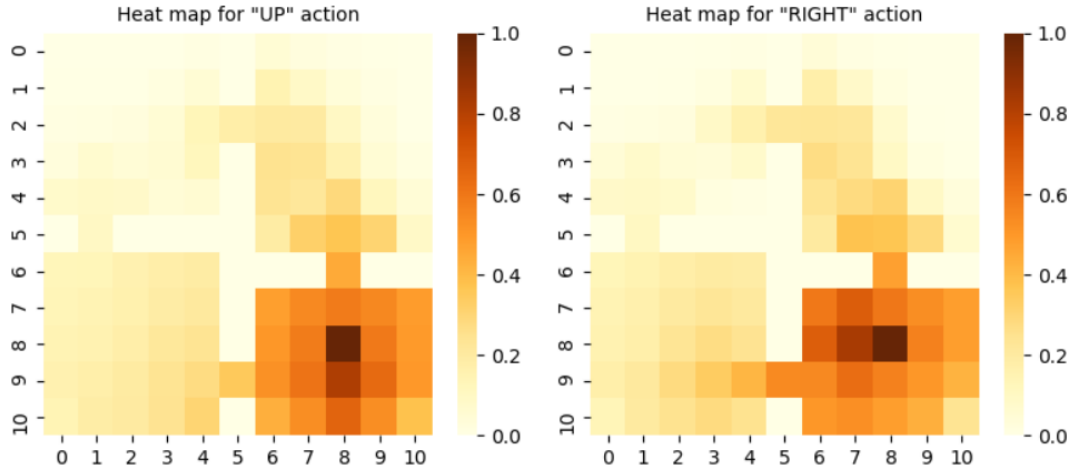

Figure 10: Heat map showing Q-values for "UP" and "RIGHT" actions for goal G2 with initial state in center of room 4

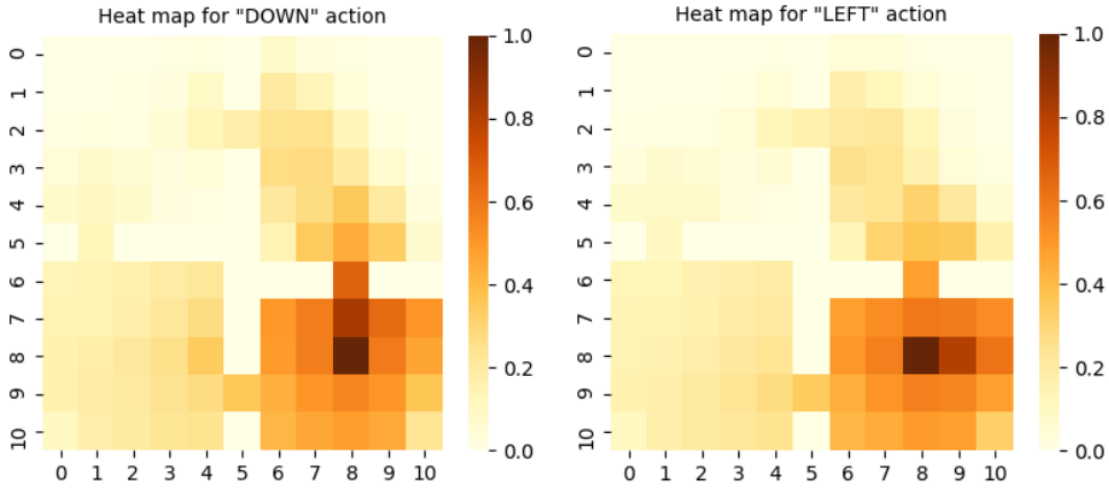

Figure 11: Heat map showing Q-values for "DOWN" and "LEFT" actions for goal G2 with initial state in center of room 4

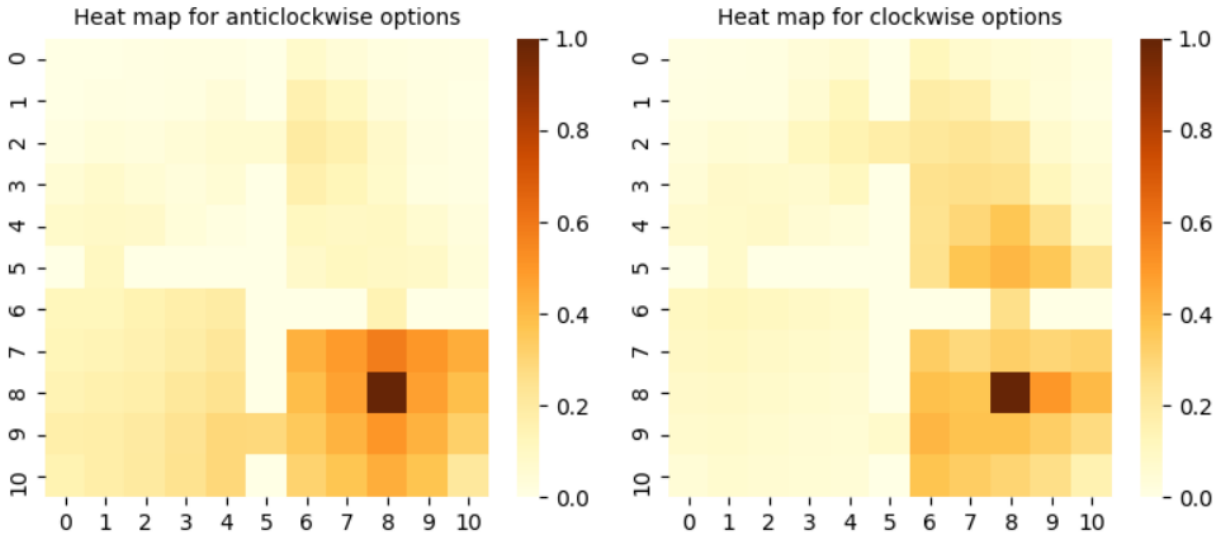

Figure 12: Heat map showing Q-values for anticlockwise and clockwise options for goal G2 with initial state in center of room 4

2 Intra-option Q-learning (Bonus)

In this section we implement the intra-option Q-learning, where the agent learns simultaneously many options in off-policy and select the optimal one. To learn the optimal options policy (reaching terminal state in less number of steps) we used a step reward of -1 (used for that option only and this reward is not used to solve the original problem).

We plot the heat maps for the learned Q-values for both goal states G1 and G2 in the following figures.

Heat maps for goal G1:

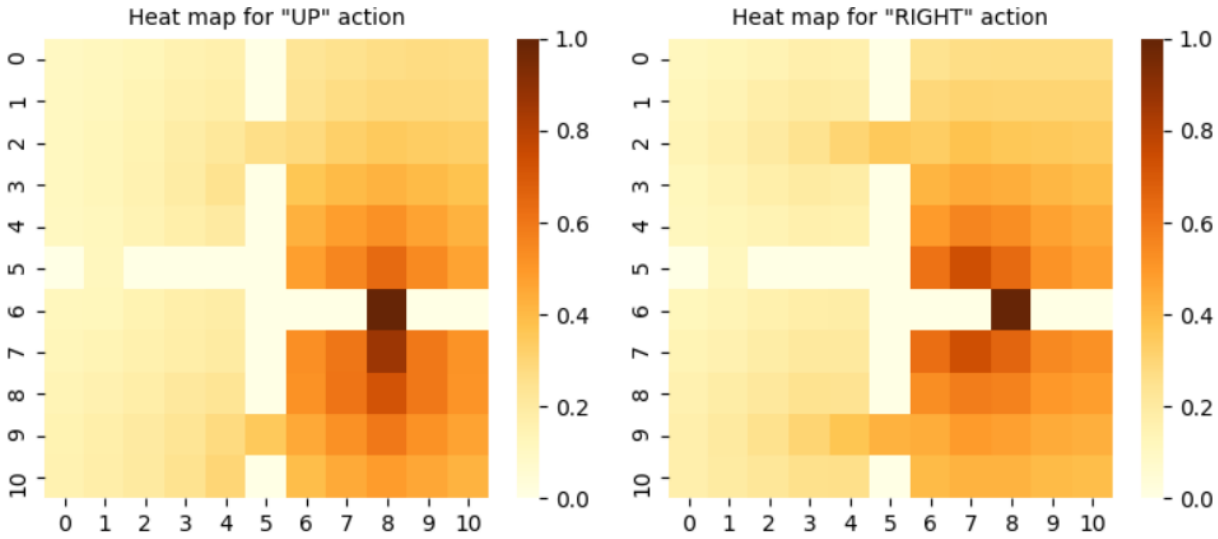

Figure 13: Heat map showing Q-values for "UP" and "RIGHT" actions for goal G1

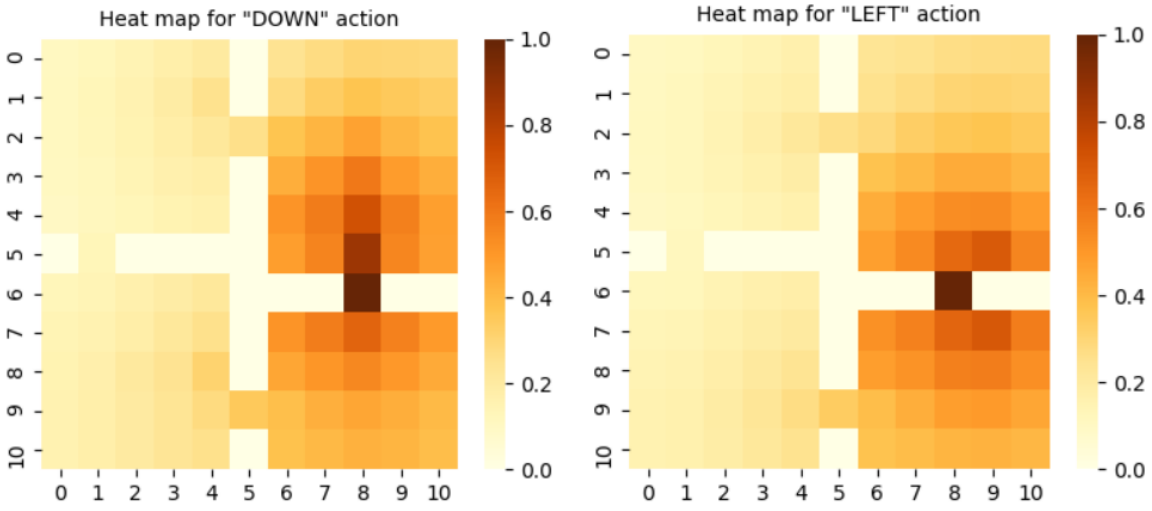

Figure 14: Heat map showing Q-values for "DOWN" and "LEFT" actions for goal G1

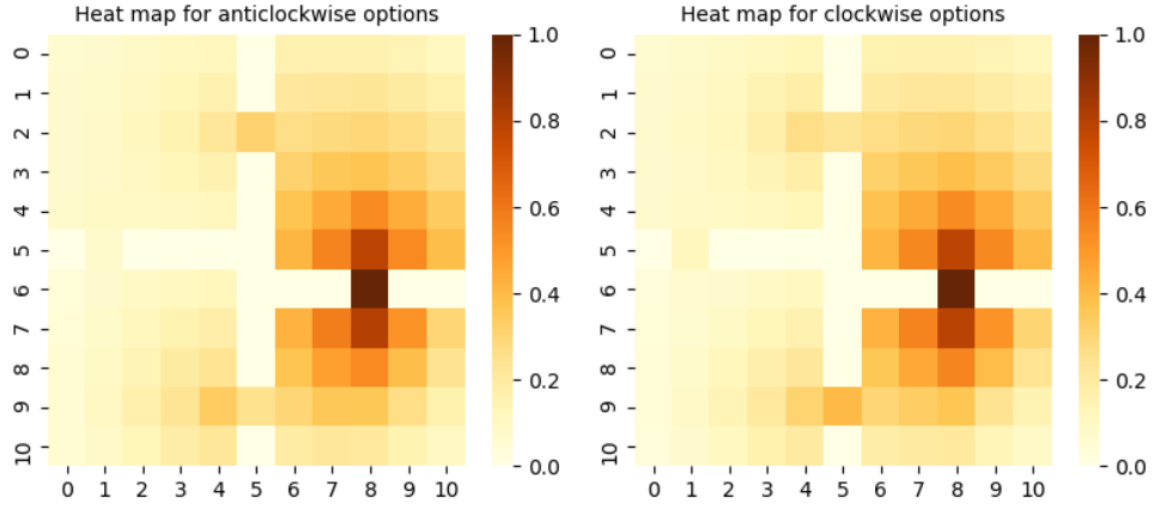

Figure 15: Heat map showing Q-values for anticlockwise and clockwise options for goal G1

Heat maps for goal G2:

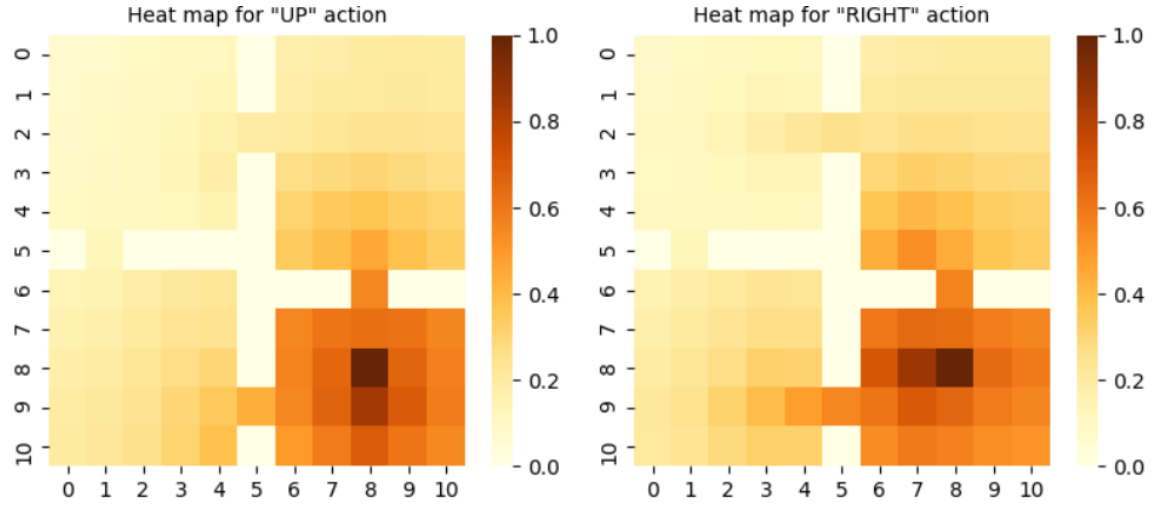

Figure 16: Heat map showing Q-values for "UP" and "RIGHT" actions for goal G2

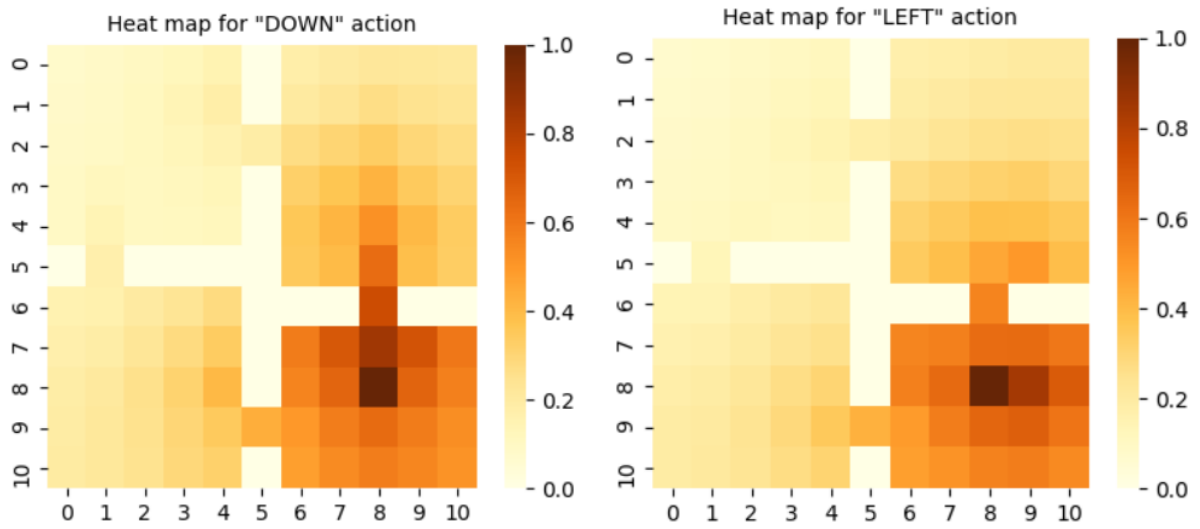

Figure 17: Heat map showing Q-values for "DOWN" and "LEFT" actions for goal G2

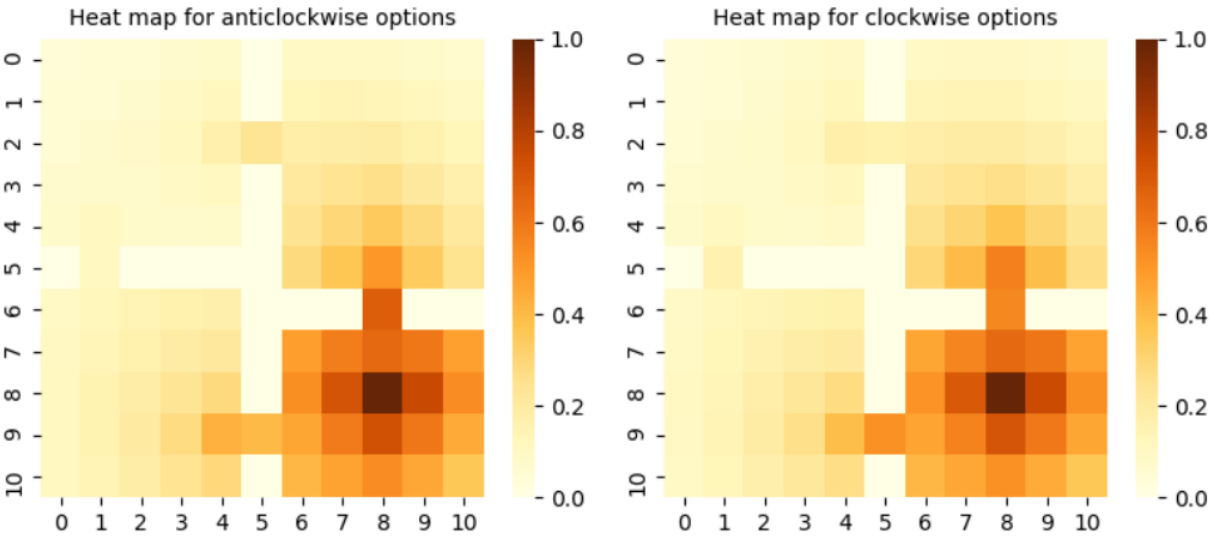

Figure 18: Heat map showing Q-values for anticlockwise and clockwise options for goal G2

2.1 Observations (for the bonus question)

(i) From the heat maps of the learned Q-values for intra-option Q-learning, we notice some similar observations as for SMDP Q-learning such as goal state having the highest Q-values and states Q-values depends on its location relative to the goal and also to the option taken there.

(ii) But one notable thing is for same goal the states have higher Q-values in intra-option Q-learning as compared to SMDP for the multi-step options, which is evident as in intra-option we leaned and compared many similar options. For this reason it also takes much time as compared to SMDP to learn the policy.

(iii) **Initial state in center of room 4:**
If the initial state of the agent is changed to the center of room 4 (state 90), we see similar observations as in the case of SMDP Q-learning, but in this case, the agent takes primitive action much more often as compared to the previous cases.