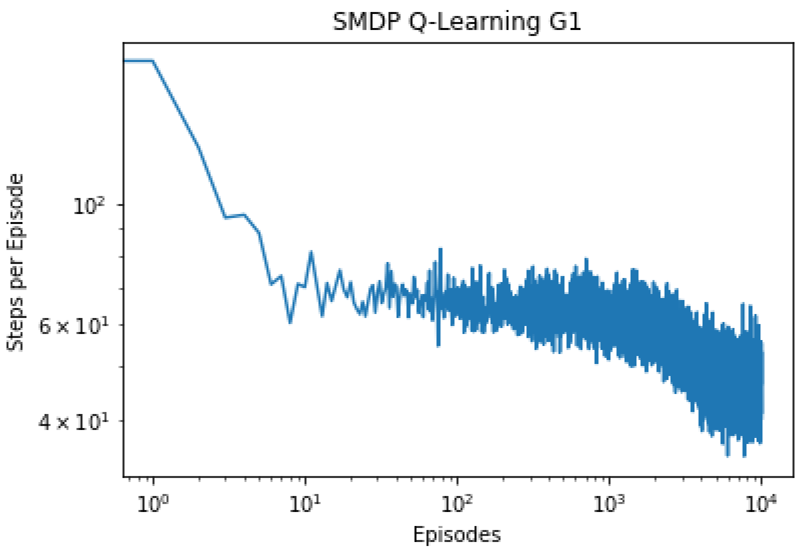
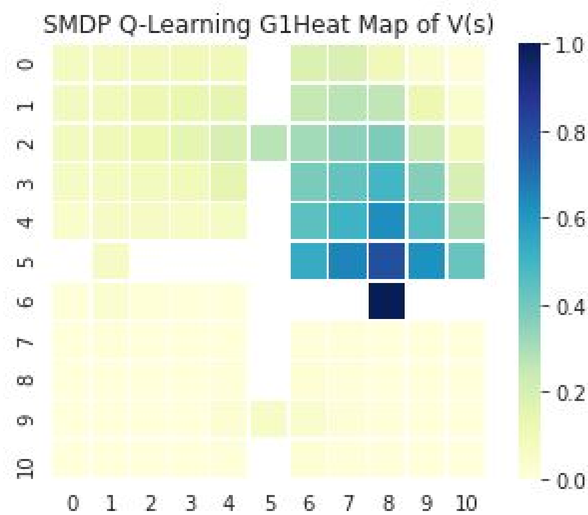
# Programming Assignment 3

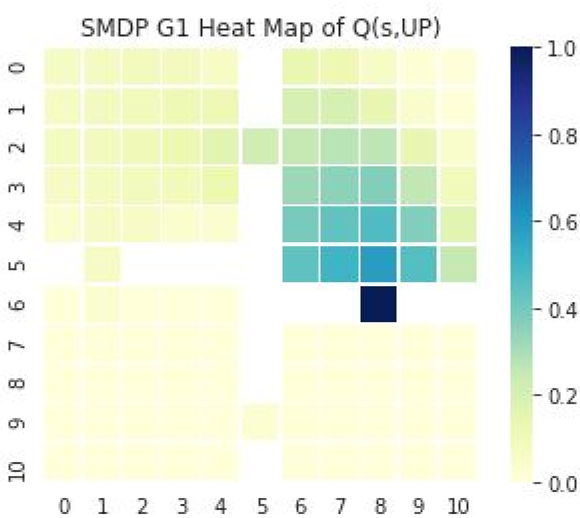
## 1. Hierarchical Reinforcement Learning

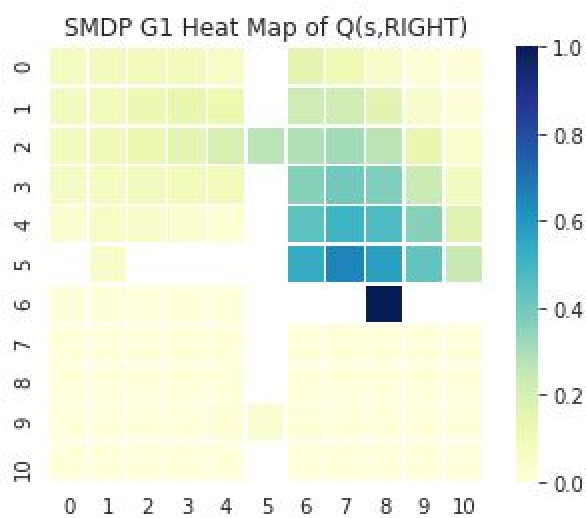
### SMDP Q - Learning

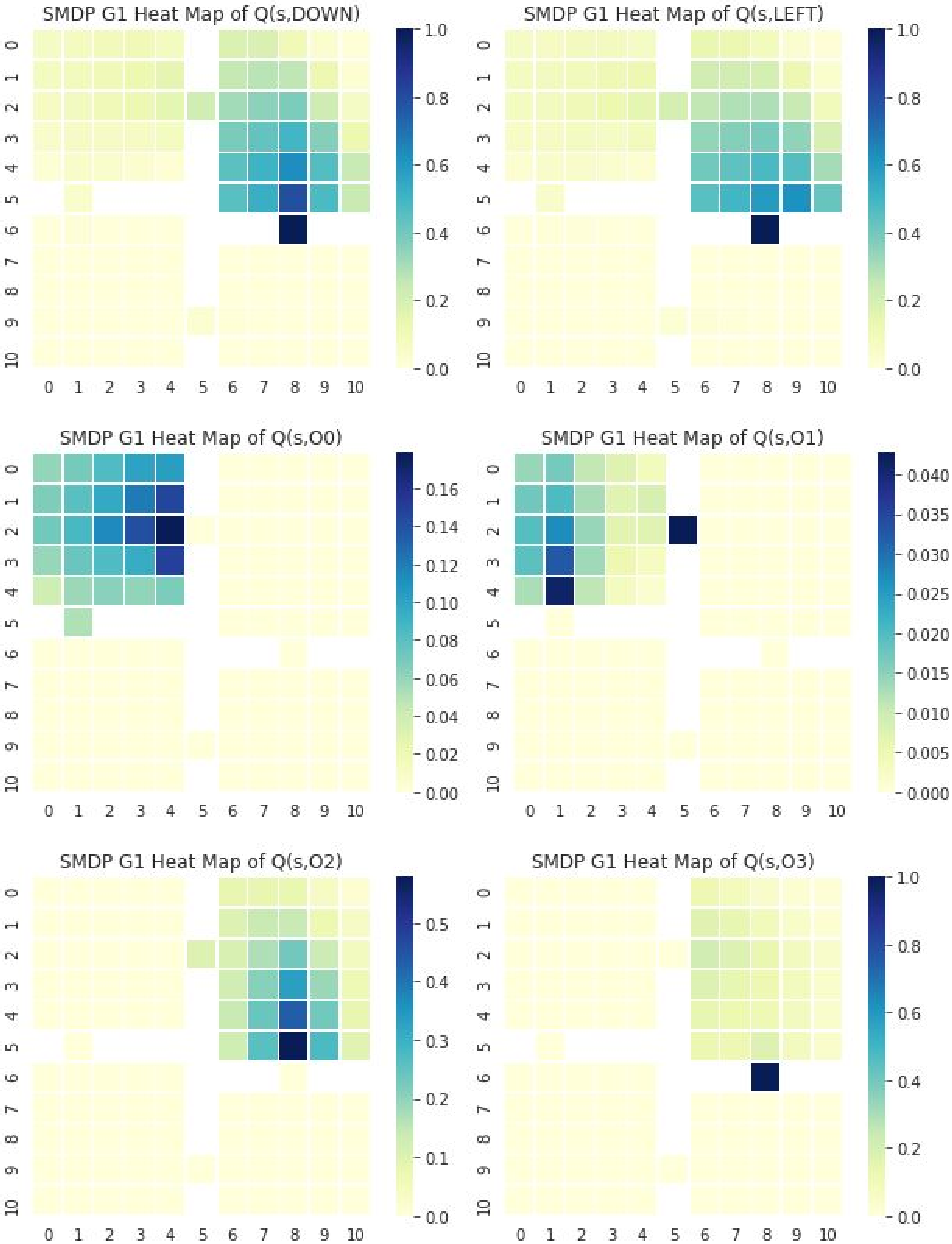
1. The steps per episode plots are similar to that of plots given in Sutton SMDP temporal abstraction paper
2. The variance at the end of these graphs is normal because these are logarithmic plots i.e. 1 to 10, 100 to 1000 takes the same space

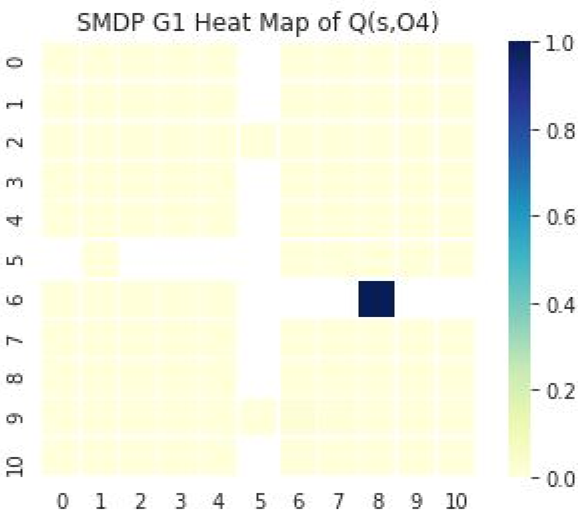


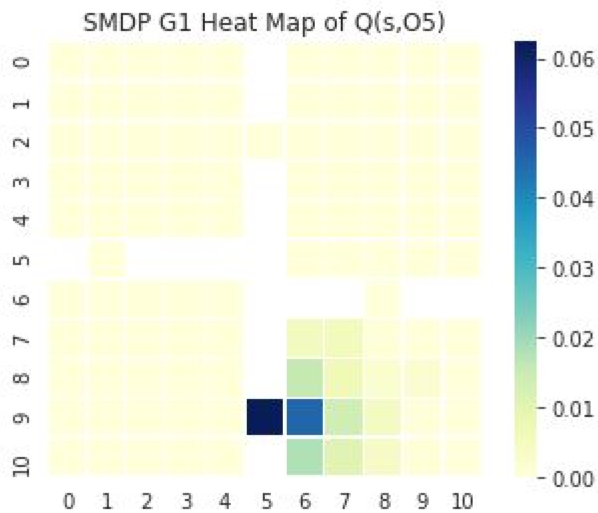



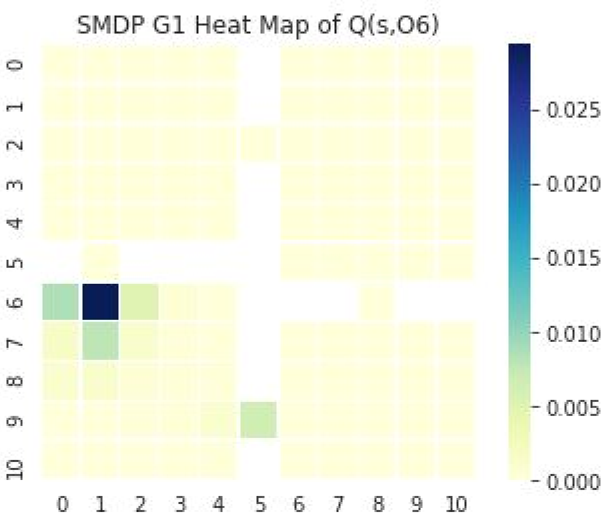



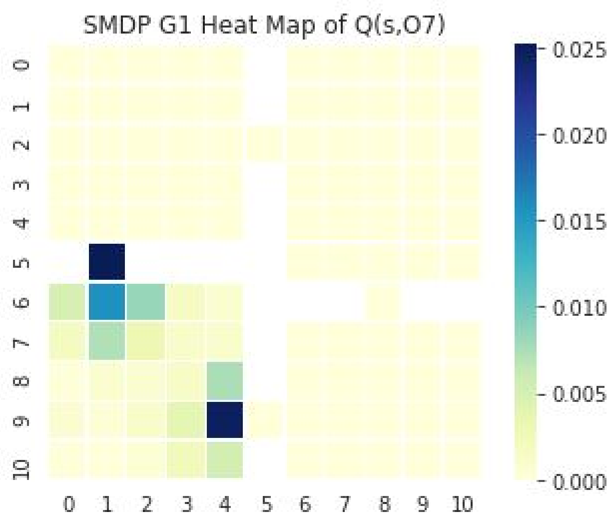



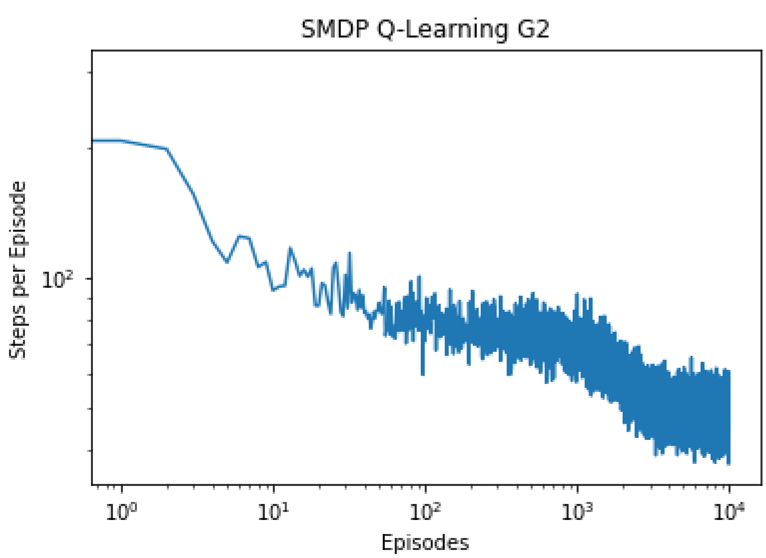

SMDP G1 Heat Map of Q(s,DOWN)
SMDP G1 Heat Map of Q(s,LEFT)
SMDP G1 Heat Map of Q(s,O0)
SMDP G1 Heat Map of Q(s,O1)
SMDP G1 Heat Map of Q(s,O2)
SMDP G1 Heat Map of Q(s,O3)

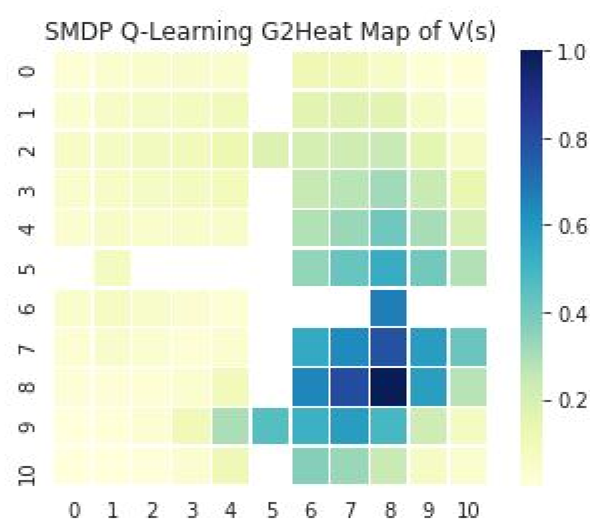

SMDP G1 Heat Map of Q(s,O4)
0 1 2 3 4 5 6 7 8 9 10
1.0
0.8
0.6
0.4
0.2
0.0

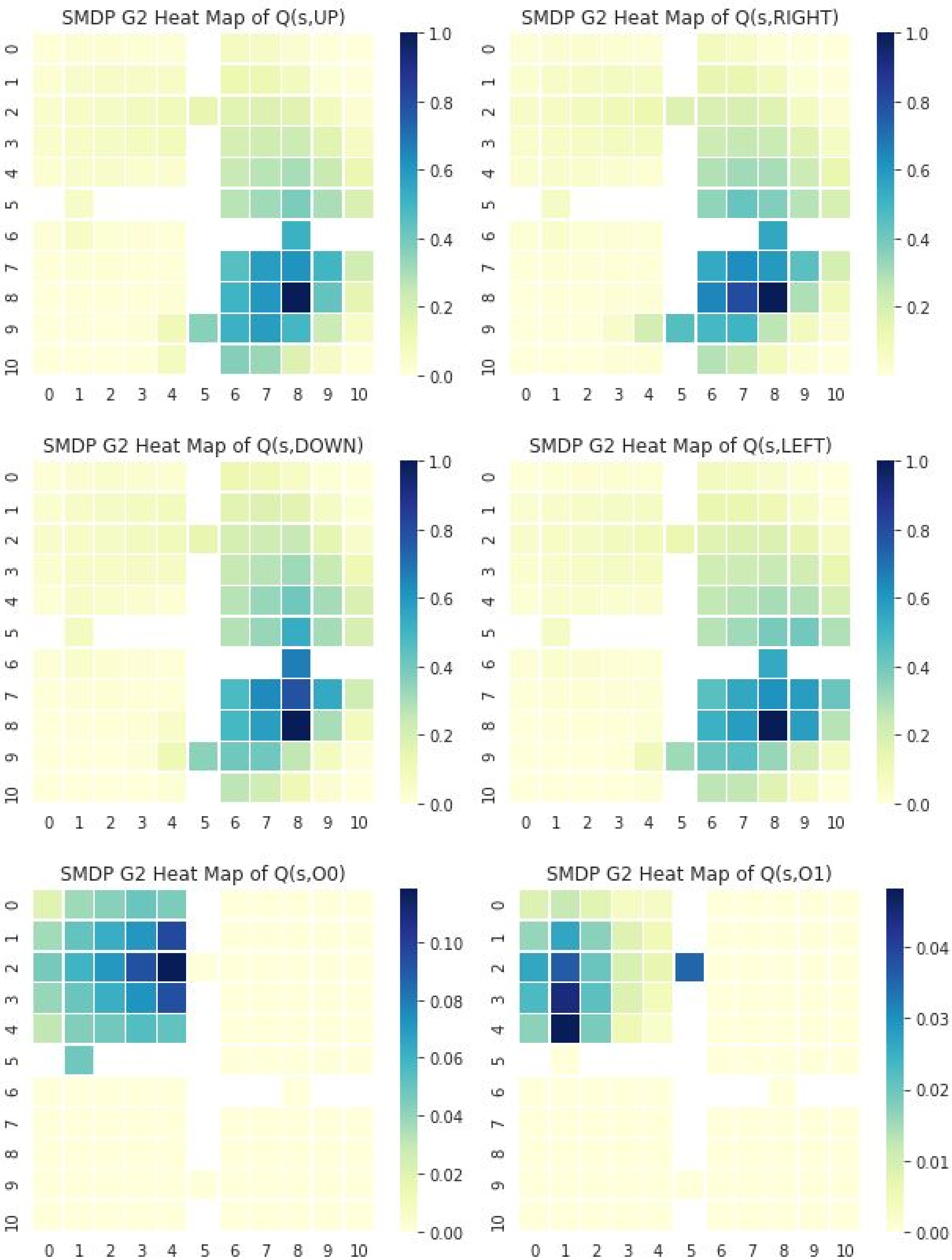

SMDP G1 Heat Map of Q(s,O5)
0 1 2 3 4 5 6 7 8 9 10
0.06
0.05
0.04
0.03
0.02
0.01
0.00

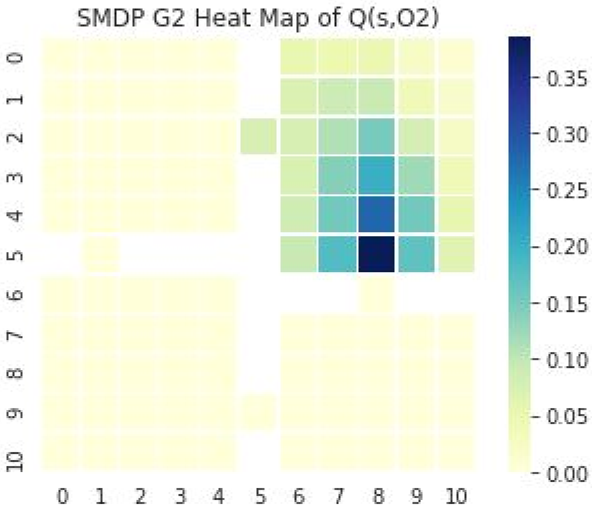

SMDP G1 Heat Map of Q(s,O6)
0 1 2 3 4 5 6 7 8 9 10
0.025
0.020
0.015
0.010
0.005
0.000

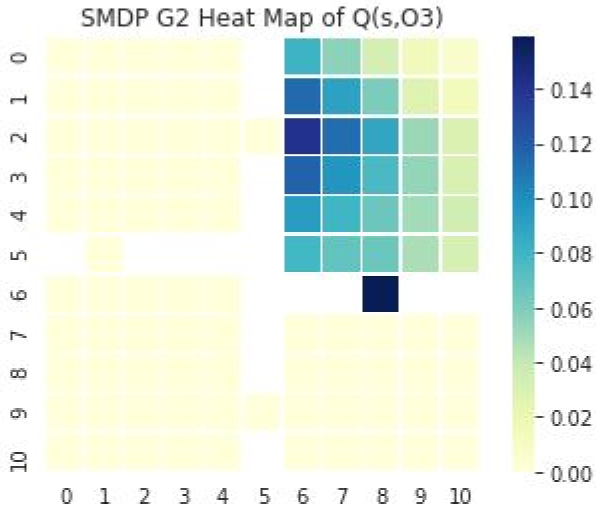

SMDP G1 Heat Map of Q(s,O7)
0 1 2 3 4 5 6 7 8 9 10
0.025
0.020
0.015
0.010
0.005
0.000

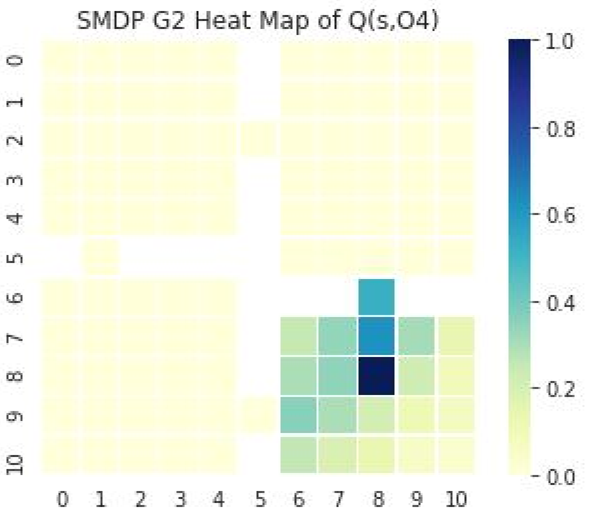

SMDP Q-Learning G2
Steps per Episode
$10^2$
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Episodes

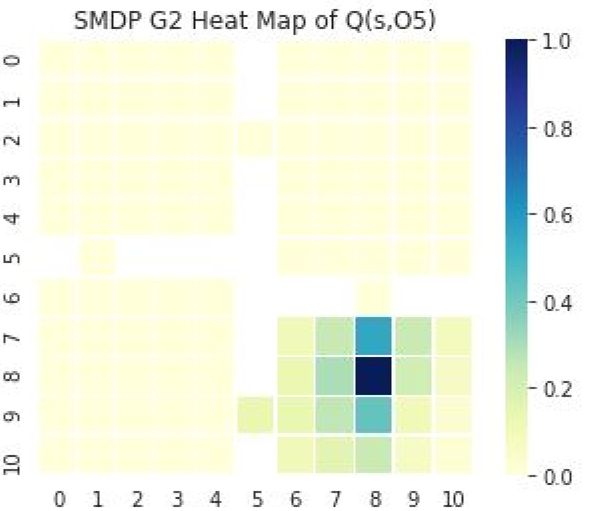

SMDP Q-Learning G2Heat Map of V(s)
0 1 2 3 4 5 6 7 8 9 10
1.0
0.8
0.6
0.4
0.2

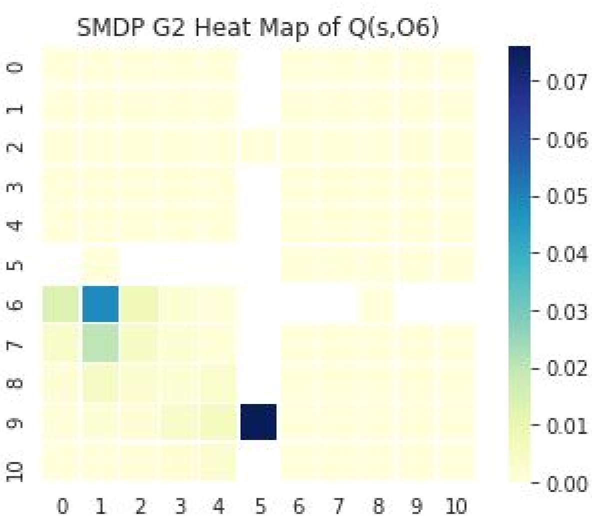

SMDP G2 Heat Map of Q(s,UP)
SMDP G2 Heat Map of Q(s,RIGHT)
SMDP G2 Heat Map of Q(s,DOWN)
SMDP G2 Heat Map of Q(s,LEFT)
SMDP G2 Heat Map of Q(s,O0)
SMDP G2 Heat Map of Q(s,O1)

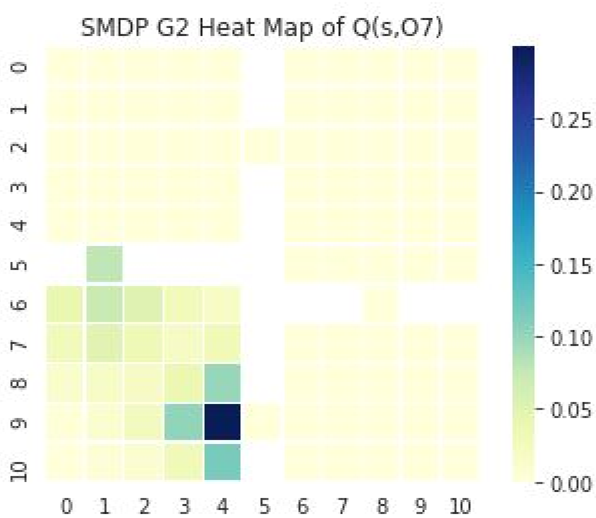

SMDP G2 Heat Map of Q(s,O2)
0 1 2 3 4 5 6 7 8 9 10
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35


SMDP G2 Heat Map of Q(s,O3)
0 1 2 3 4 5 6 7 8 9 10
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14

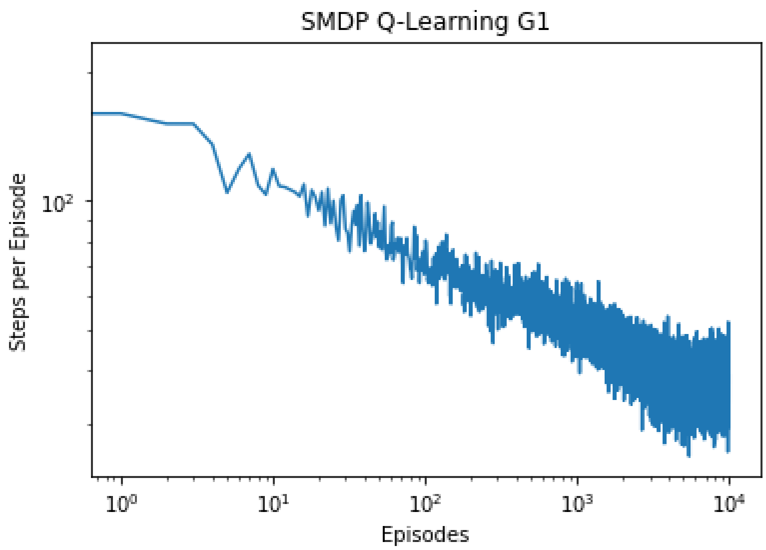

SMDP G2 Heat Map of Q(s,O4)
0 1 2 3 4 5 6 7 8 9 10
0.0 0.2 0.4 0.6 0.8 1.0

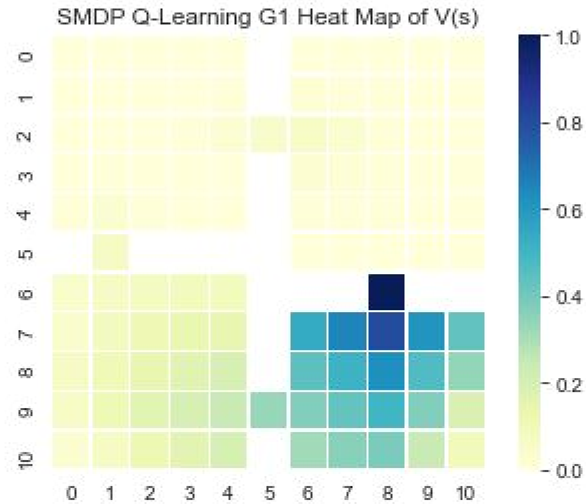

SMDP G2 Heat Map of Q(s,O5)
0 1 2 3 4 5 6 7 8 9 10
0.0 0.2 0.4 0.6 0.8 1.0

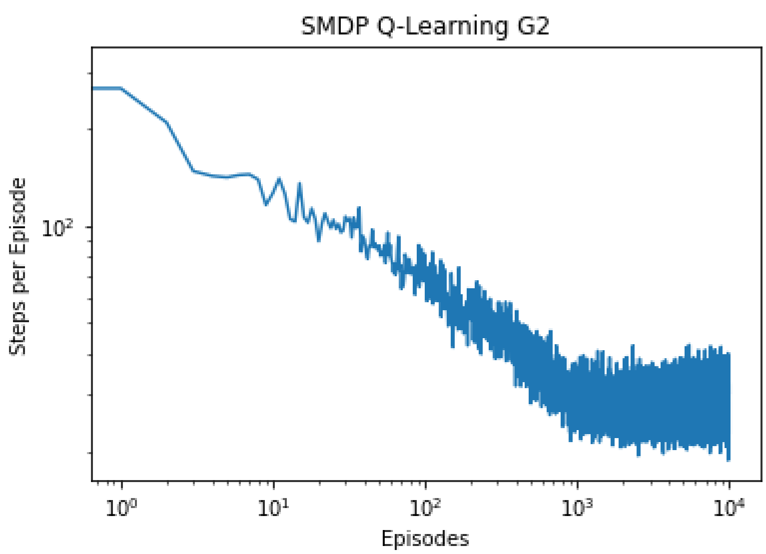

SMDP G2 Heat Map of Q(s,O6)
0 1 2 3 4 5 6 7 8 9 10
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07

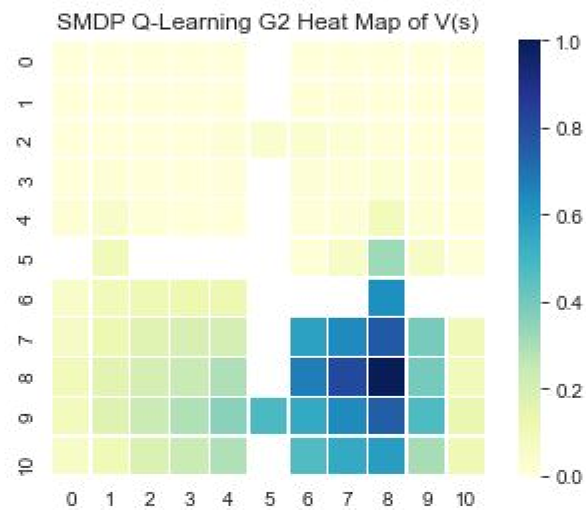

SMDP G2 Heat Map of Q(s,O7)
0 1 2 3 4 5 6 7 8 9 10
0.00 0.05 0.10 0.15 0.20 0.25

## Changing Initial State to Center of Room 4



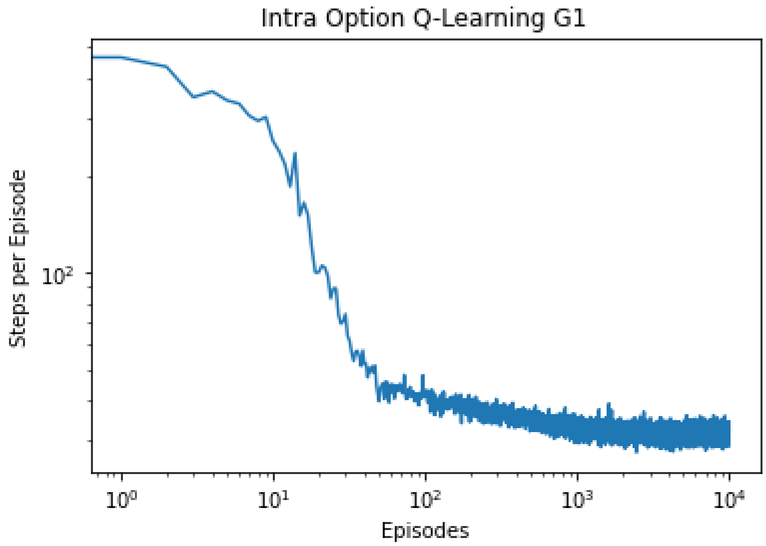



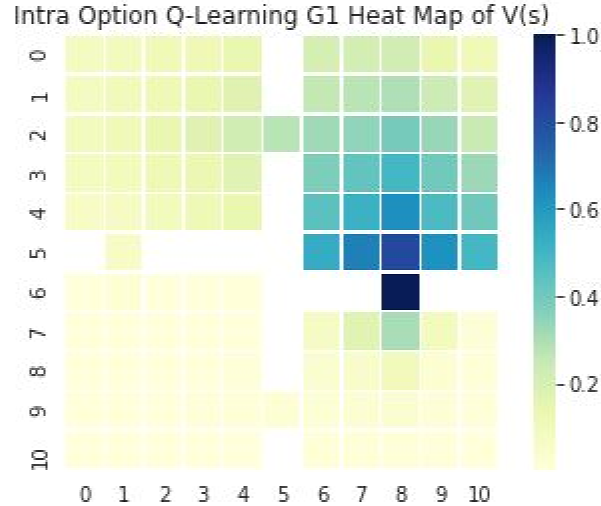



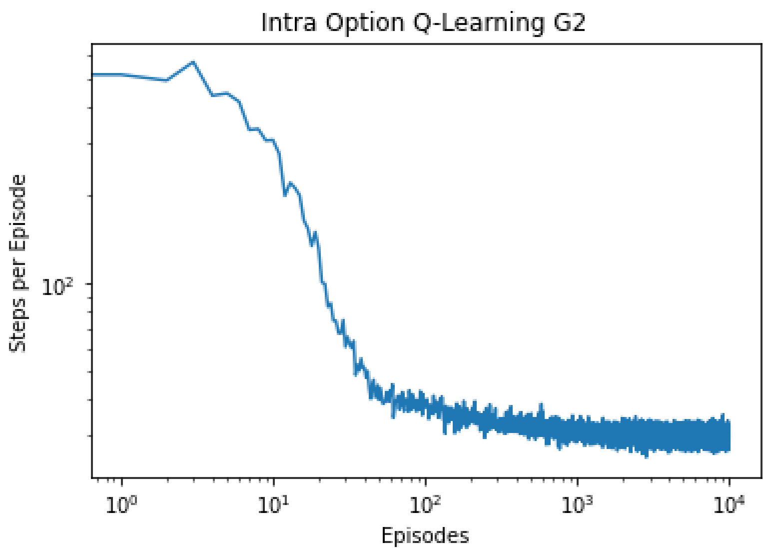

1. Training process got faster as usually the start is some random state in room 1 but now it is fixed to the center of room 4
2. The Q plots for this are at the end of the doc

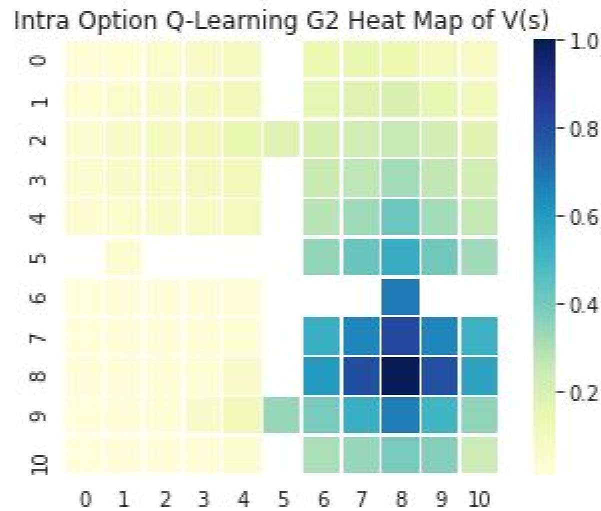
## Intra-option Q learning

1. The convergence is faster, we can see the same in the graph
2. The main reason is that in SMDP we execute option to termination before we update learn from it but here we learn from each step
3. Another reason is that we can update more than one options from single instance if the policy of different options match at a state
4. The Q plots for Intra option Q learning are at the end of the doc

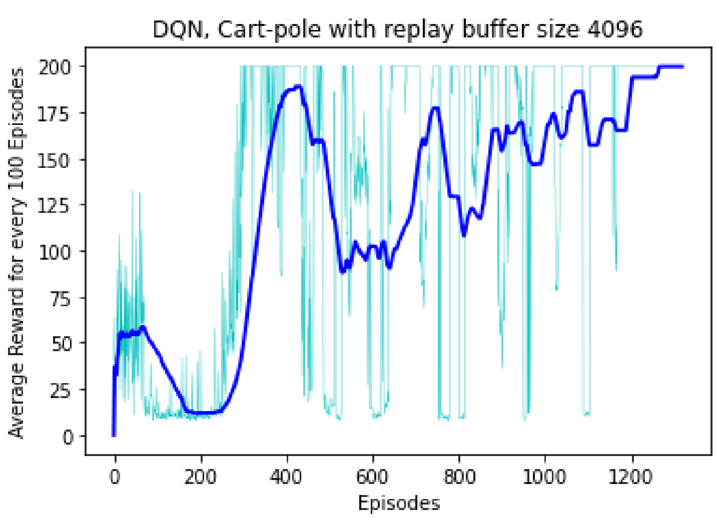


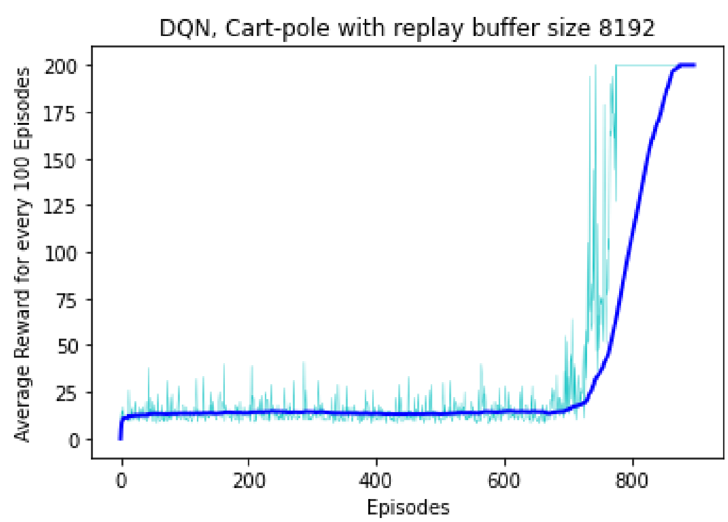



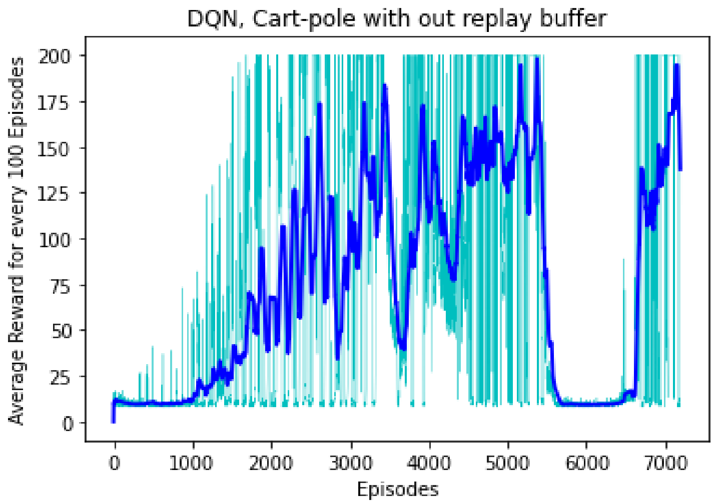



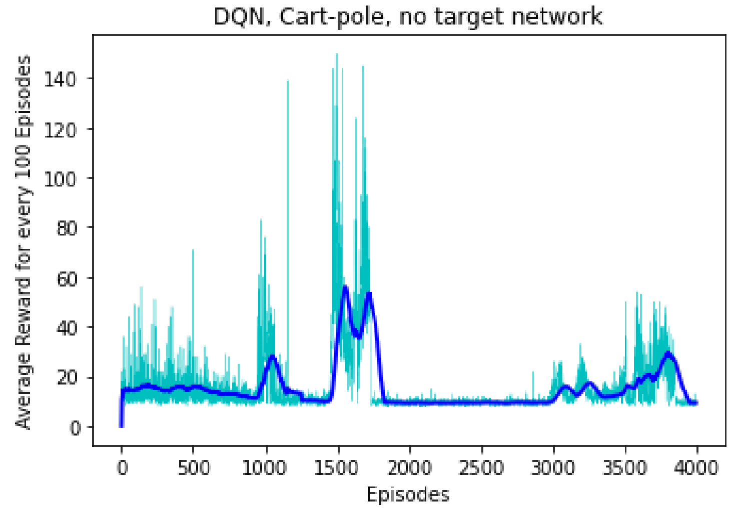




## 2. (Not so) Deep RL





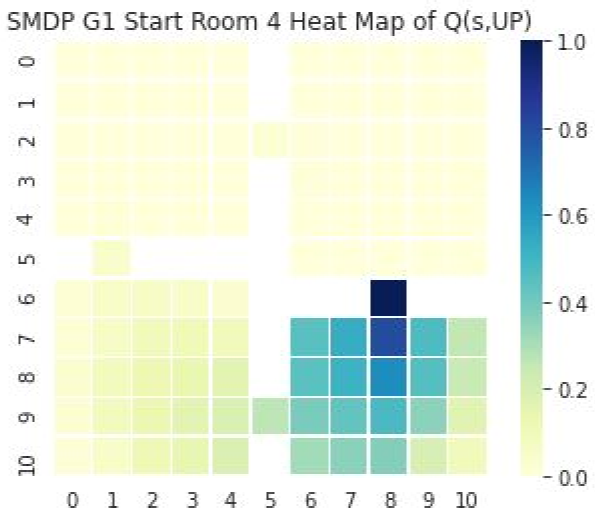

## Tuned Hyperparameters

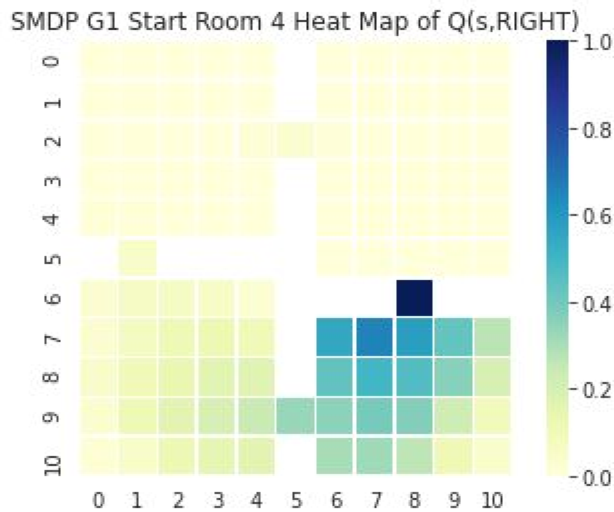
1. REPLAY_MEMORY_SIZE = 4096, 8192
2. EPSILON = 0.5
3. EPSILON_DECAY = 0.999, 0.9999
4. HIDDEN1_SIZE = 32
5. HIDDEN2_SIZE = 16
6. EPISODES_NUM = 1300
7. MAX_STEPS = 200
8. LEARNING_RATE = 0.001
9. MINIBATCH_SIZE = 16 , 64
10. DISCOUNT_FACTOR = 0.999
11. TARGET_UPDATE_FREQ = 50 , 100
12. EPSILON_MIN = 0.02

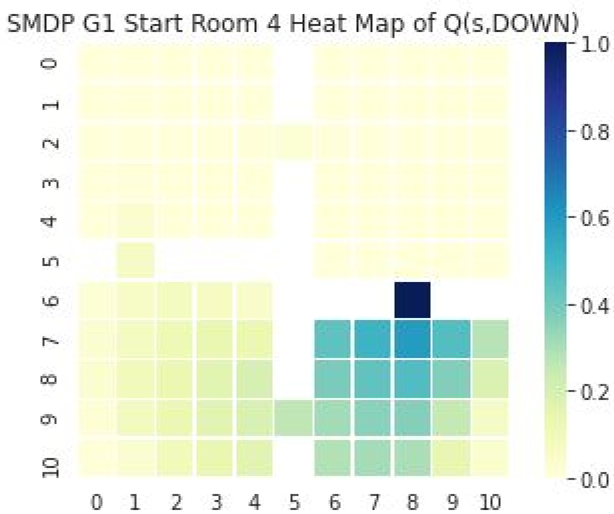
## Observations

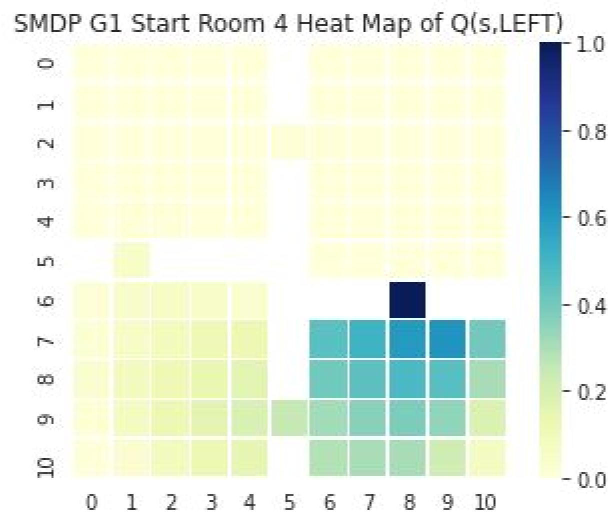
1. The second set of parameters are giving the best results we can see the convergence in the graphs, below are some observations
2. On decreasing target update freq the variance of model increasing and on increasing convergence is taking longer
3. Low discount factors unable to clearly differentiate the increase in steps after some threshold
4. Increasing mini-batch helping to achieve convergence faster as minibatch can be computed parallelly with more than one workers
5. Replay memory size as we know helps in reducing variance due to adjacent transitions and also to create uniform randomness in picking transition
6. We can see that in the 3rd plot the high variance, the more episodes owe to the fact that we do only one update in a step i.e. the mini-batch size will automatically be 1 if we remove the replay buffer.
7. Increasing the learning rate the steps are oscillating between 100-200 and on decreasing they couldn't reach 200 so it took time to tune it. This looks similar to that of stopping at local minima in a typical gradient ascent method
8. We can see that with no target there is no convergence at all. The reason for that is we are chasing a nonstationary target, we are bootstrapping. Neural networks aren't efficient in doing this, so to stabilize this variance we use target network and update for every n steps.

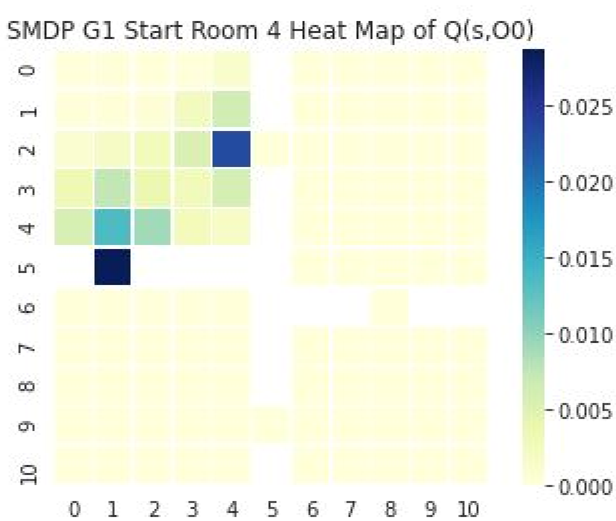
Q plots for start state in room no. 4

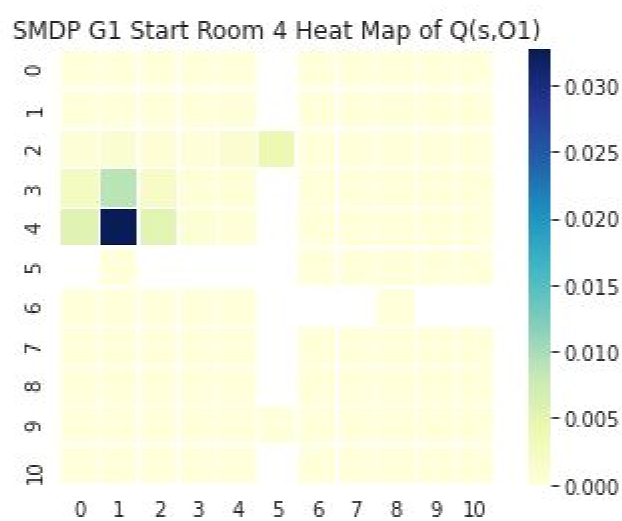


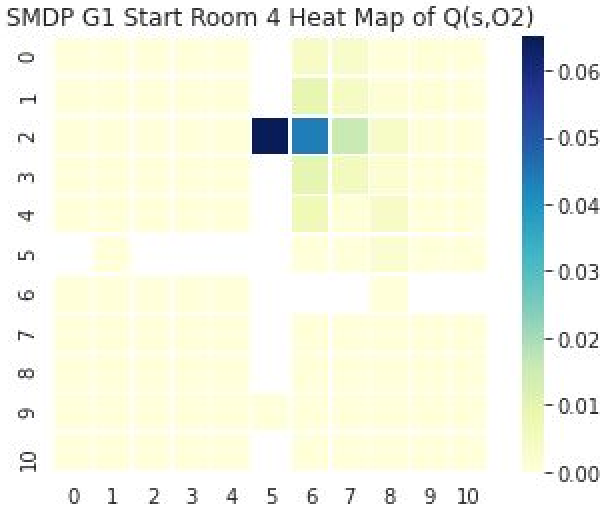



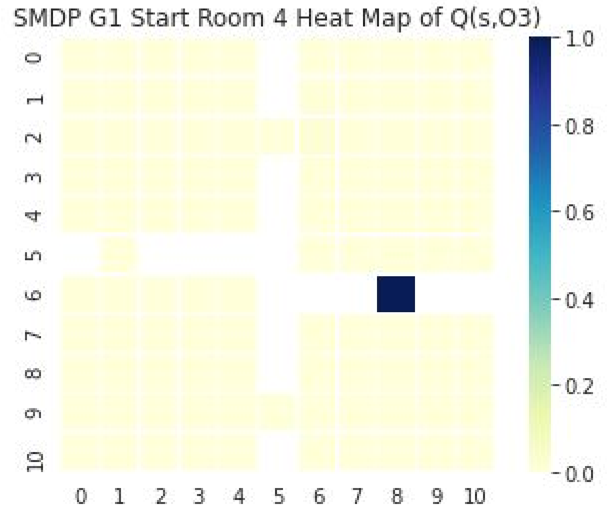



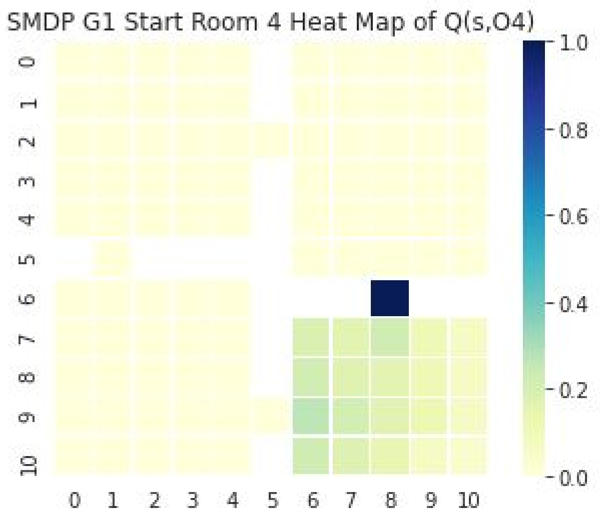



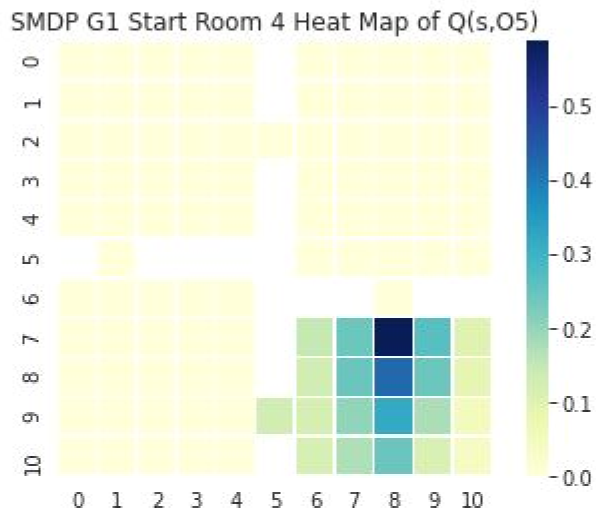



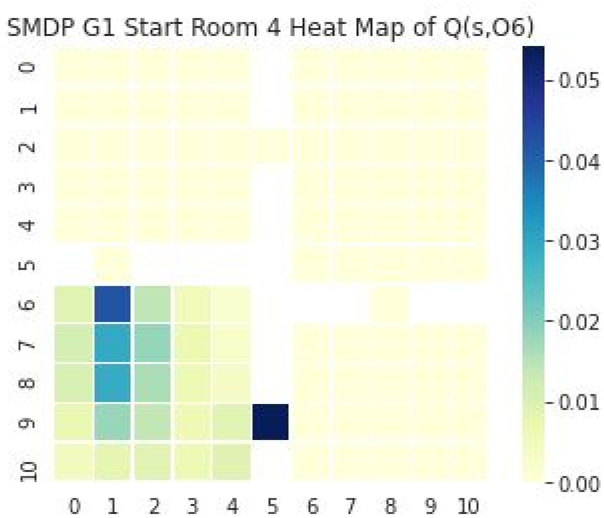



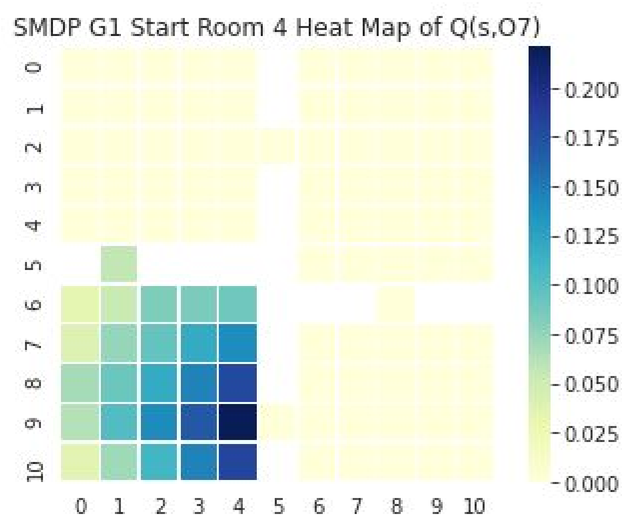

SMDP G1 Start Room 4 Heat Map of Q(s,O2)
0.06
0.05
0.04
0.03
0.02
0.01
0.00


SMDP G1 Start Room 4 Heat Map of Q(s,O3)
1.0
0.8
0.6
0.4
0.2
0.0

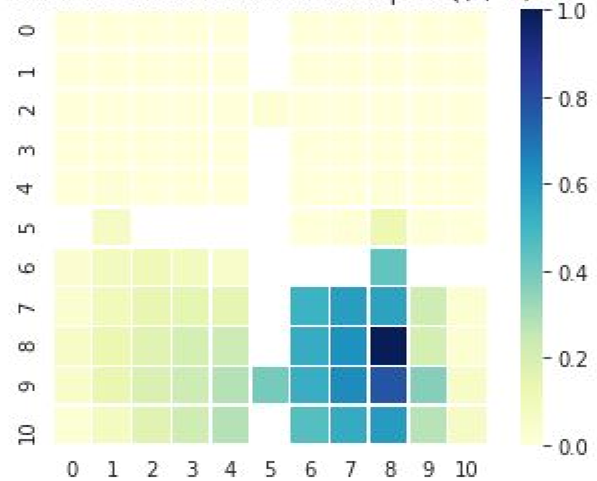

SMDP G1 Start Room 4 Heat Map of Q(s,O4)
1.0
0.8
0.6
0.4
0.2
0.0


SMDP G1 Start Room 4 Heat Map of Q(s,O5)
0.5
0.4
0.3
0.2
0.1
0.0

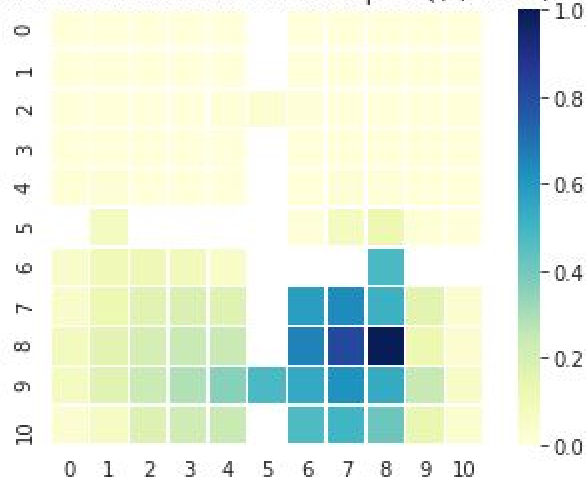

SMDP G1 Start Room 4 Heat Map of Q(s,O6)
0.05
0.04
0.03
0.02
0.01
0.00


SMDP G1 Start Room 4 Heat Map of Q(s,O7)
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000

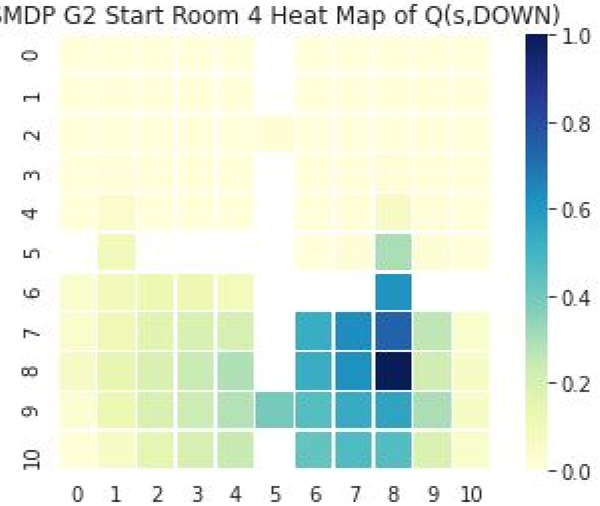
SMDP G2 Start Room 4 Heat Map of Q(s,UP)



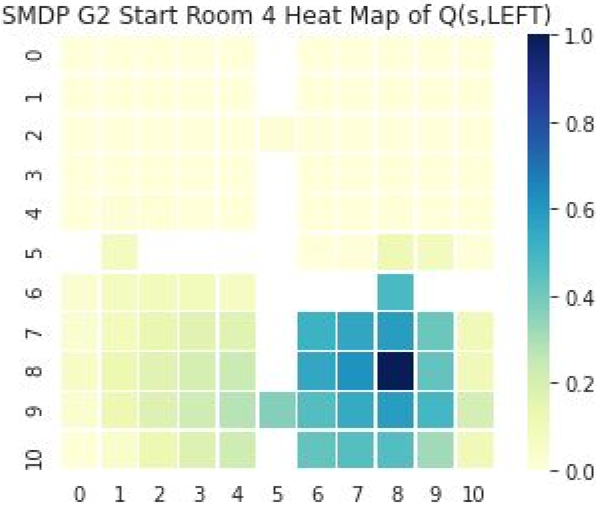

SMDP G2 Start Room 4 Heat Map of Q(s,RIGHT)



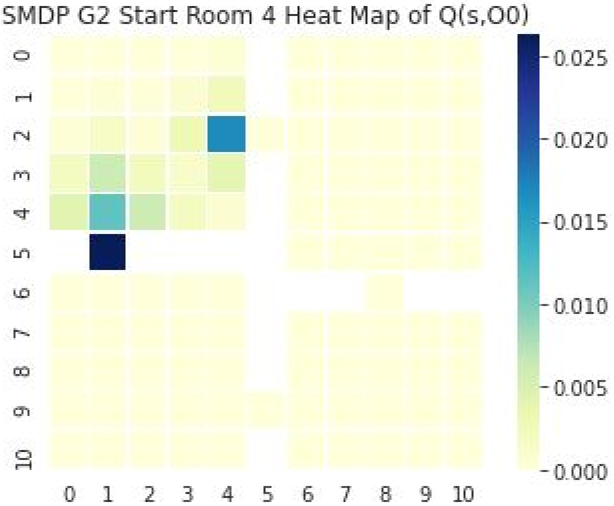

SMDP G2 Start Room 4 Heat Map of Q(s,DOWN)



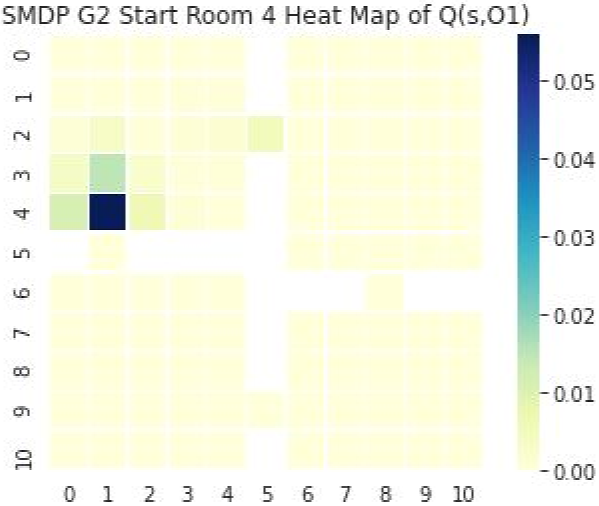

SMDP G2 Start Room 4 Heat Map of Q(s,LEFT)

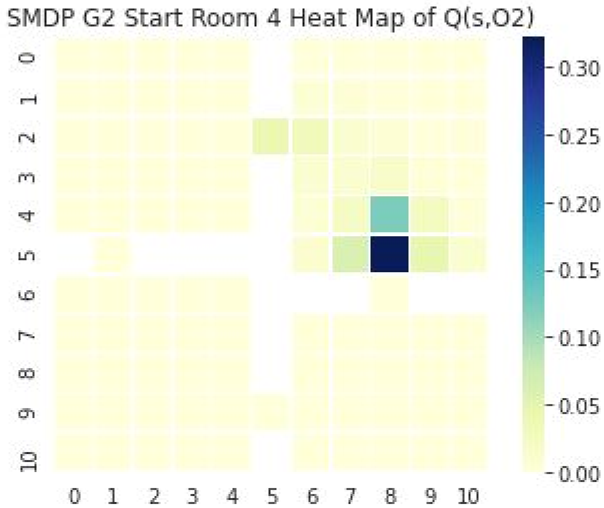


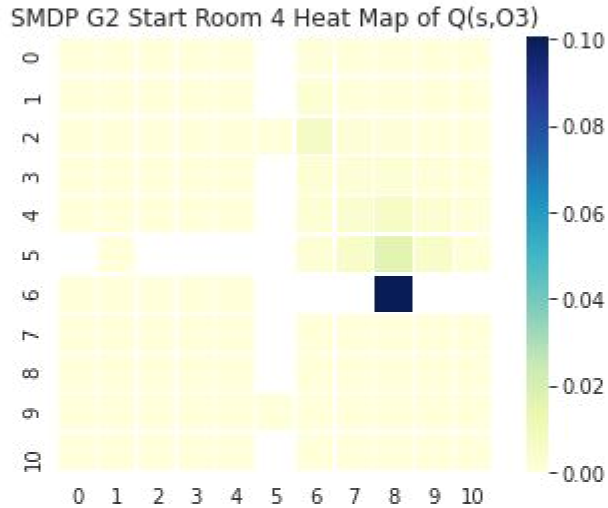

SMDP G2 Start Room 4 Heat Map of Q(s,O0)

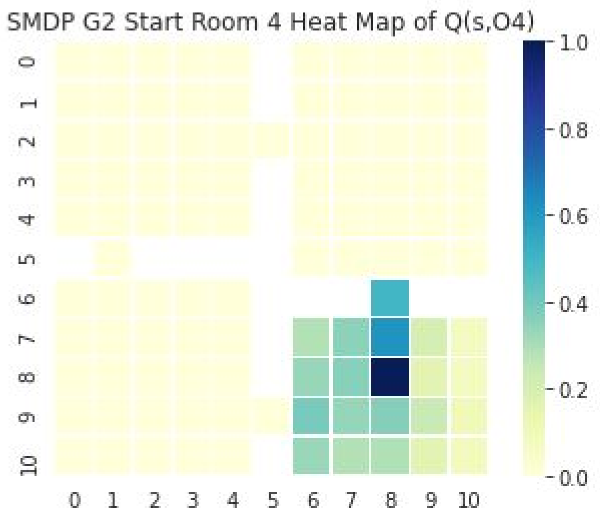


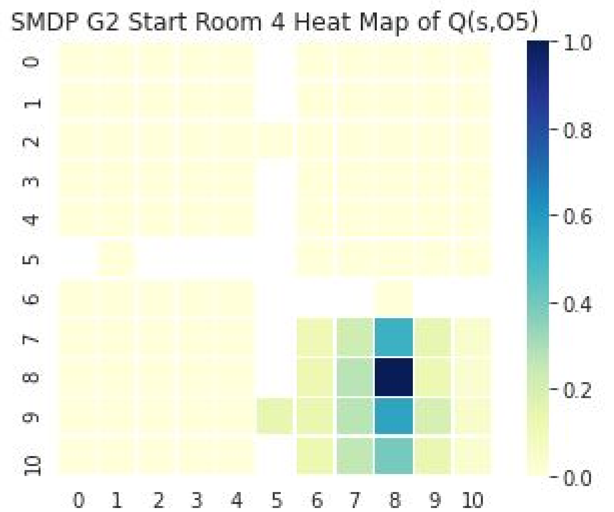

SMDP G2 Start Room 4 Heat Map of Q(s,O1)

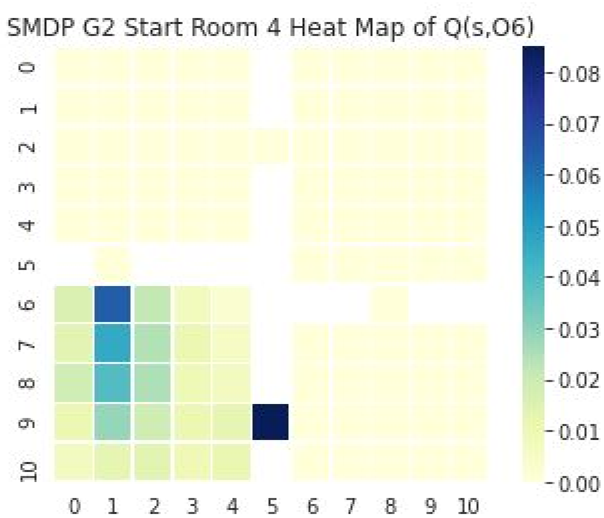


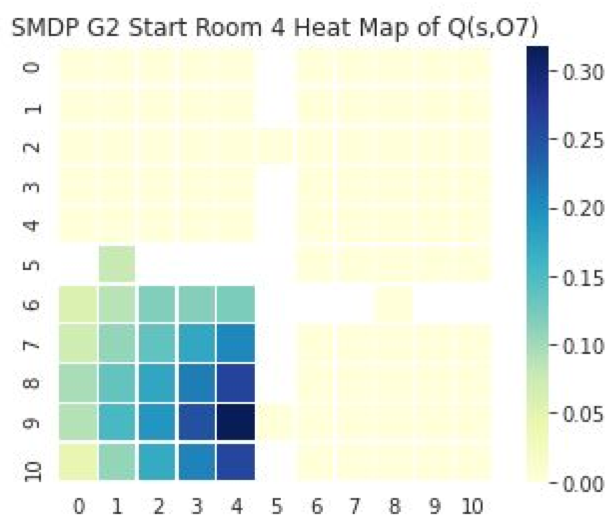

SMDP G2 Start Room 4 Heat Map of Q(s,O2)


SMDP G2 Start Room 4 Heat Map of Q(s,O3)

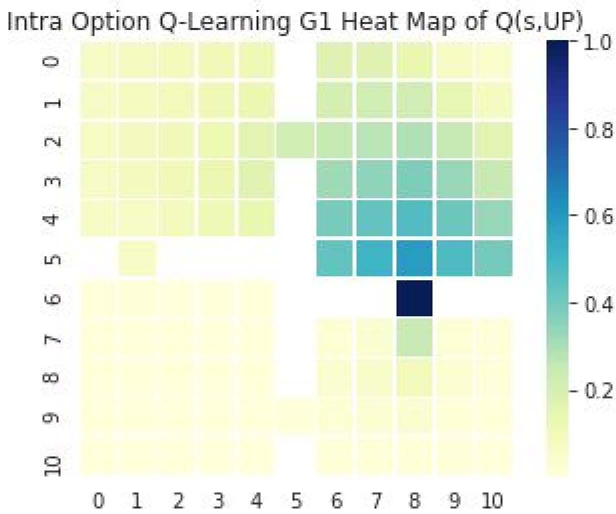

SMDP G2 Start Room 4 Heat Map of Q(s,O4)

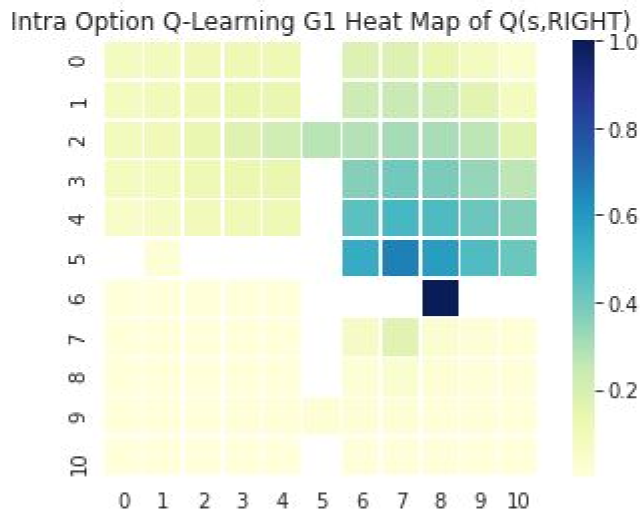

SMDP G2 Start Room 4 Heat Map of Q(s,O5)

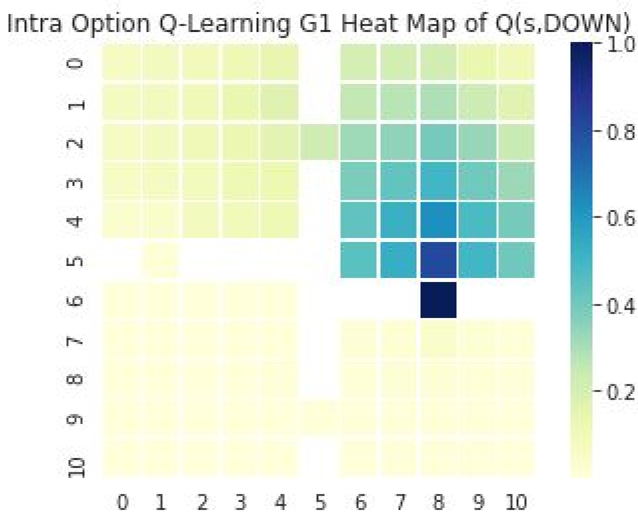

SMDP G2 Start Room 4 Heat Map of Q(s,O6)

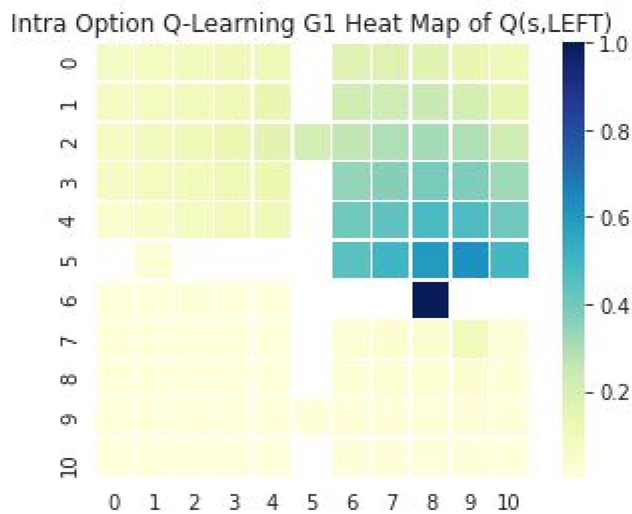

SMDP G2 Start Room 4 Heat Map of Q(s,O7)

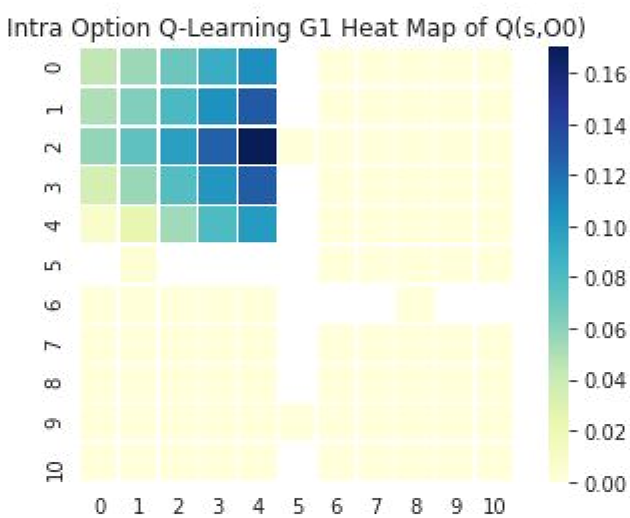
## Q plots for Intra-option Q-learning

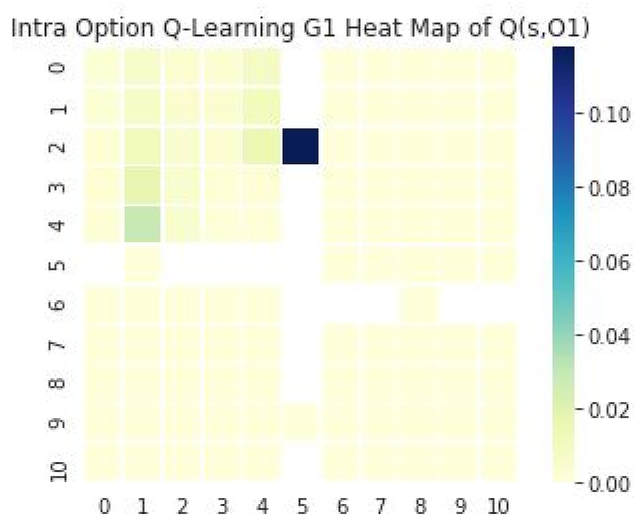


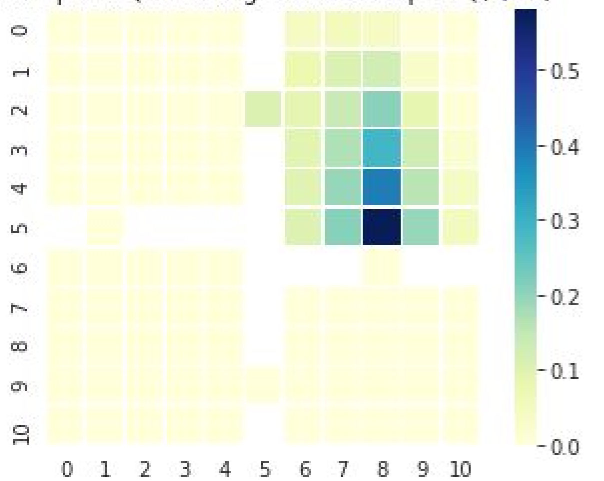



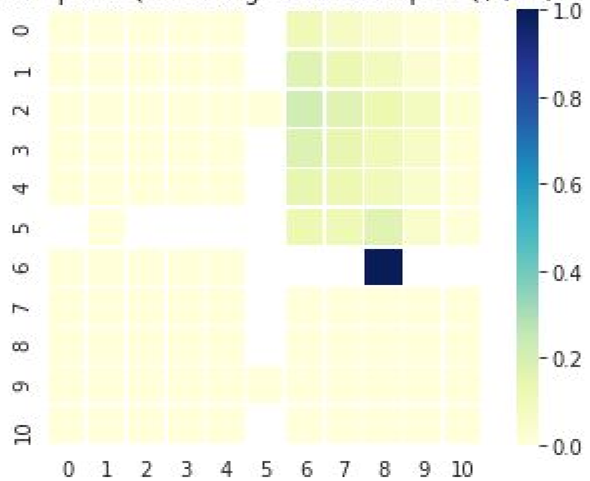



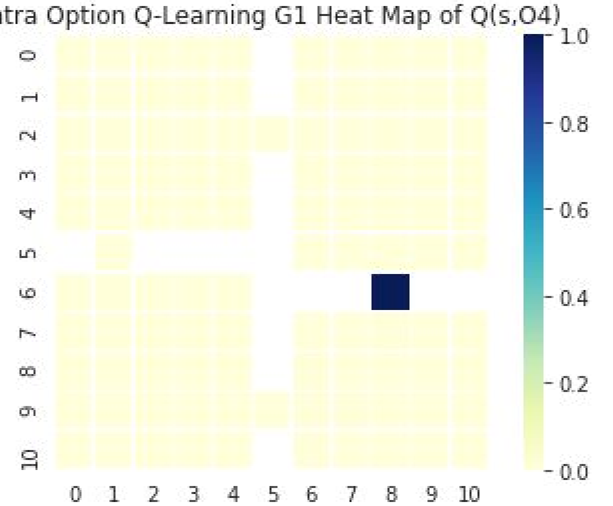
Intra Option Q-Learning G1 Heat Map of Q(s,O2)



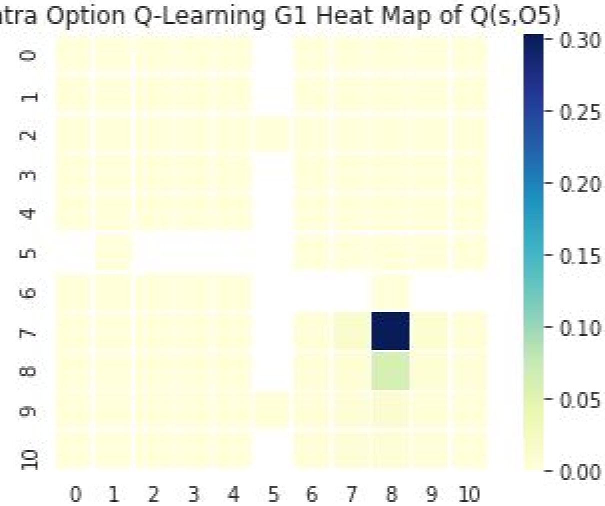
Intra Option Q-Learning G1 Heat Map of Q(s,O3)



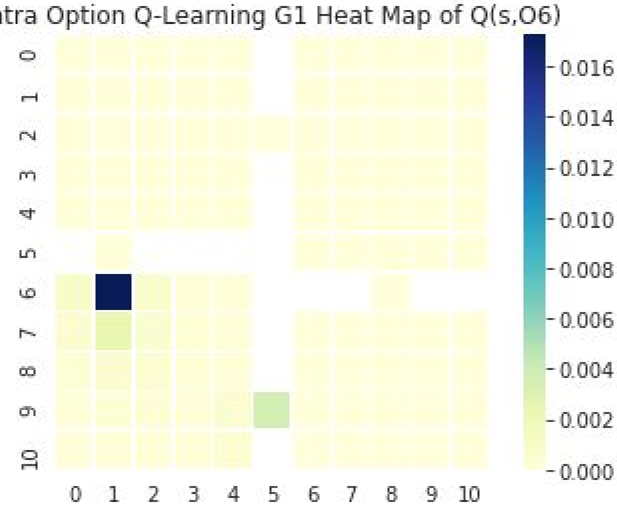
Intra Option Q-Learning G1 Heat Map of Q(s,O4)



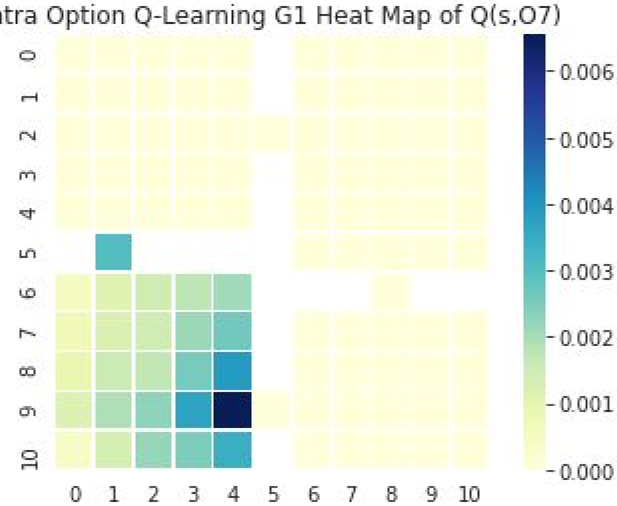
Intra Option Q-Learning G1 Heat Map of Q(s,O5)

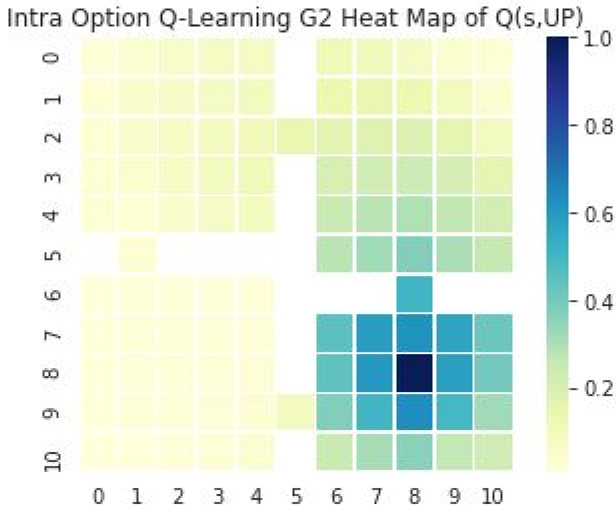


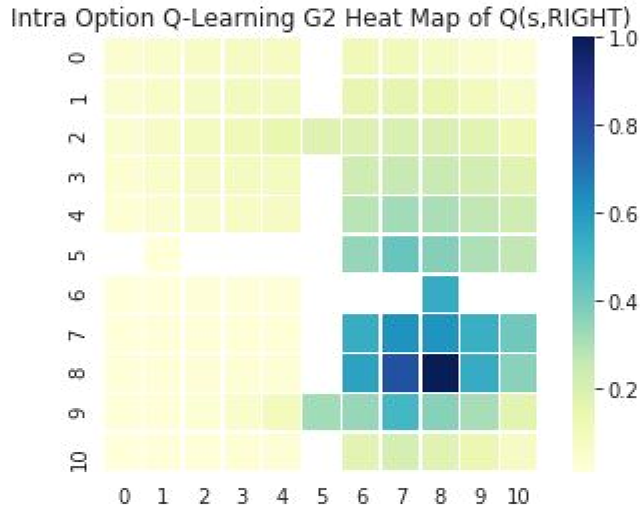
Intra Option Q-Learning G1 Heat Map of Q(s,O6)

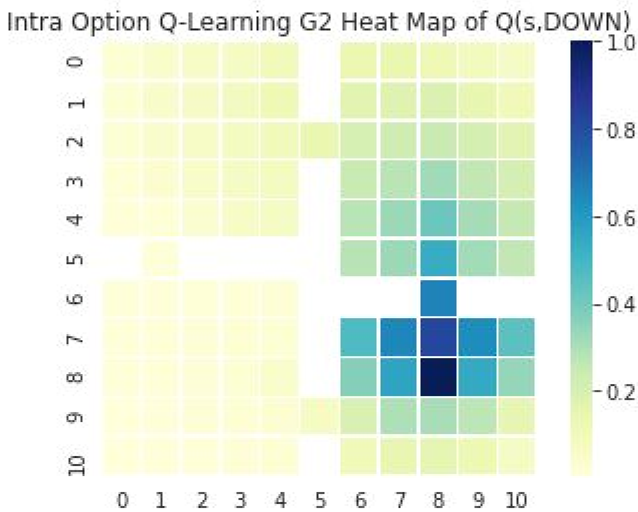


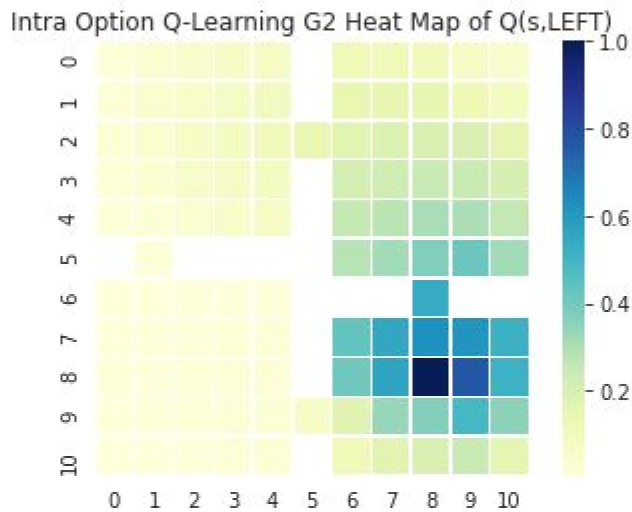
Intra Option Q-Learning G1 Heat Map of Q(s,O7)

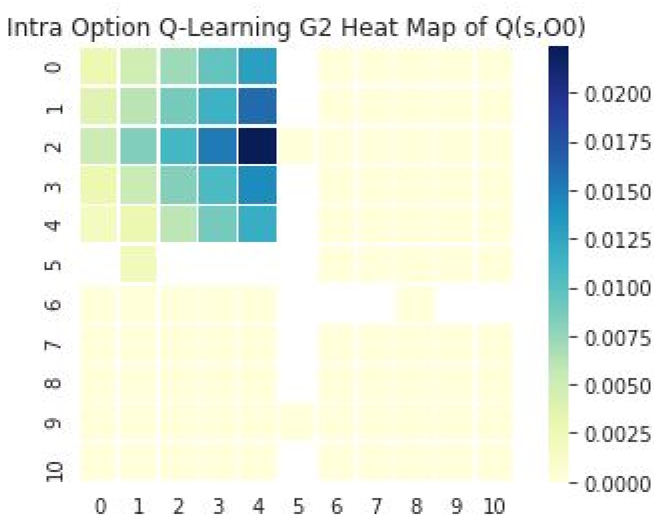


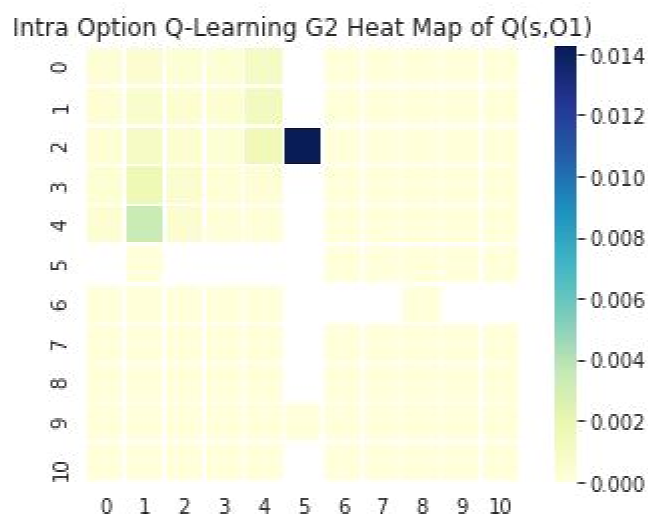

Intra Option Q-Learning G2 Heat Map of Q(s,UP)
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
1.0 0.8 0.6 0.4 0.2

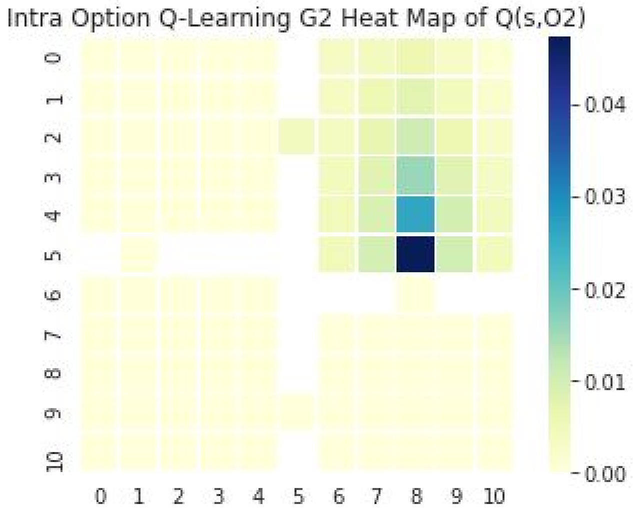

Intra Option Q-Learning G2 Heat Map of Q(s,RIGHT)
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
1.0 0.8 0.6 0.4 0.2

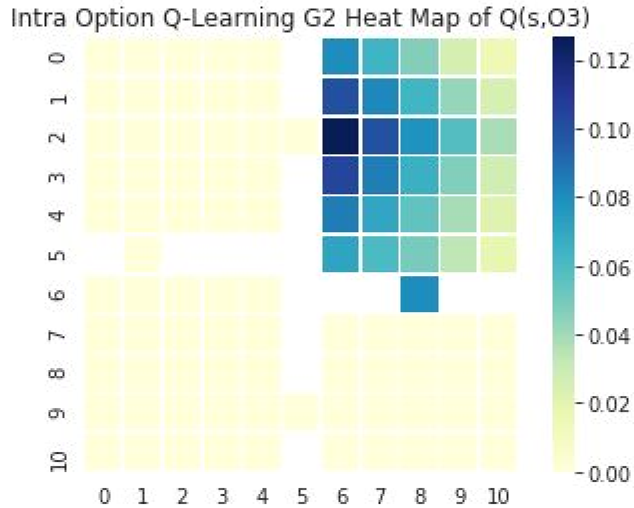

Intra Option Q-Learning G2 Heat Map of Q(s,DOWN)
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
1.0 0.8 0.6 0.4 0.2

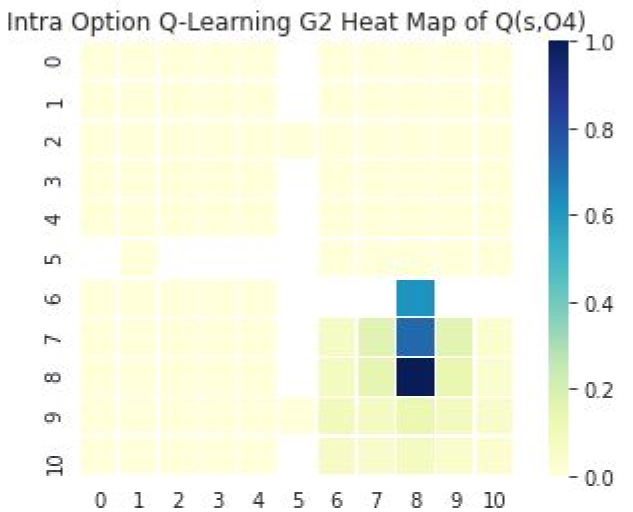

Intra Option Q-Learning G2 Heat Map of Q(s,LEFT)
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
1.0 0.8 0.6 0.4 0.2

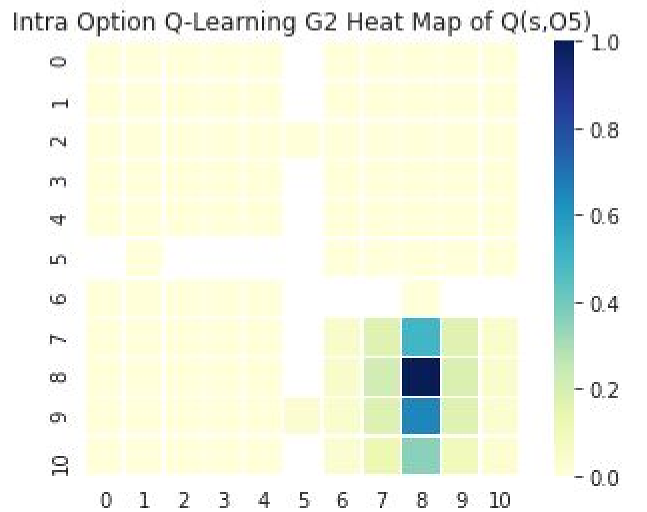

Intra Option Q-Learning G2 Heat Map of Q(s,O0)
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
0.0200 0.0175 0.0150 0.0125 0.0100 0.0075 0.0050 0.0025 0.0000

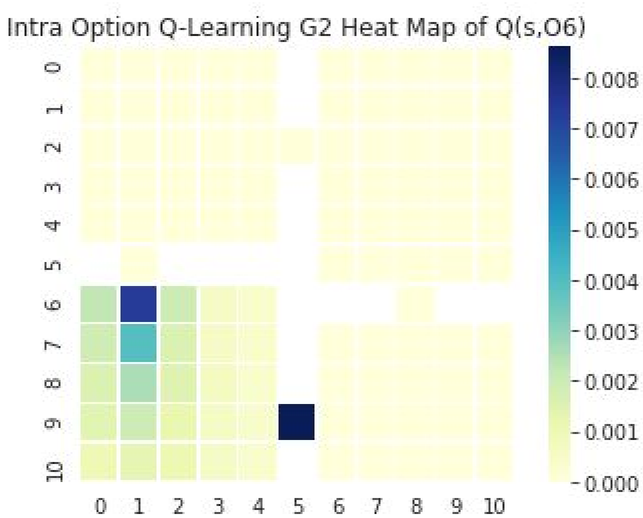

Intra Option Q-Learning G2 Heat Map of Q(s,O1)
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
0.014 0.012 0.010 0.008 0.006 0.004 0.002 0.000

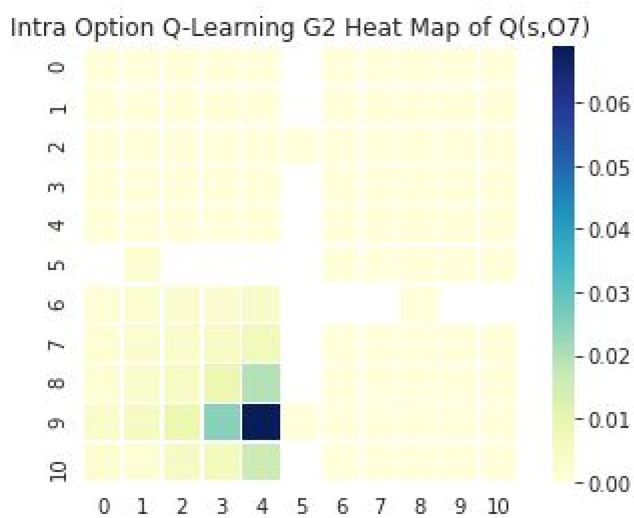

Intra Option Q-Learning G2 Heat Map of Q(s,O2)
0.04
0.03
0.02
0.01
0.00


Intra Option Q-Learning G2 Heat Map of Q(s,O3)
0.12
0.10
0.08
0.06
0.04
0.02
0.00


Intra Option Q-Learning G2 Heat Map of Q(s,O4)
1.0
0.8
0.6
0.4
0.2
0.0


Intra Option Q-Learning G2 Heat Map of Q(s,O5)
1.0
0.8
0.6
0.4
0.2
0.0


Intra Option Q-Learning G2 Heat Map of Q(s,O6)
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000


Intra Option Q-Learning G2 Heat Map of Q(s,O7)
0.06
0.05
0.04
0.03
0.02
0.01
0.00