Deep Reinforcement Learning S2

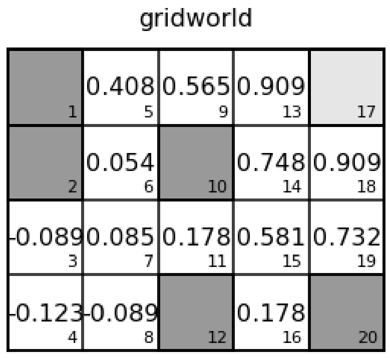
Dynamic Programming, Monte-Carlo Sarsa, Q-learning

Neta Bar-gil 305609836 & Yair Lahad 205493018

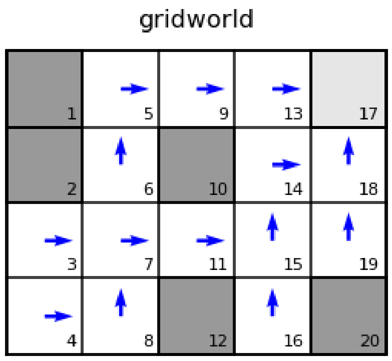
1. Dynamic Programming:

   For the Dynamic Programming optimal solution for the Frozen lake, we decided to implement Value iteration model using synchronous dynamic programming. We provide the MDP model that includes PSS' matrix for each action, and the given reward for each state. (Appendix 1) The optimal solution described in figures bellow:

Optimal state values:



Optimal policy:



2. Monte-Carlo control:

   In order to find the best approximation to the optimal solution found using the DP, we tried various parameters and compared the optimal state values from MC to the optimal state values received from the DP. Each parameter was tested for 500 episodes.

   Tested Parameters:

   Epsilon = 0.9

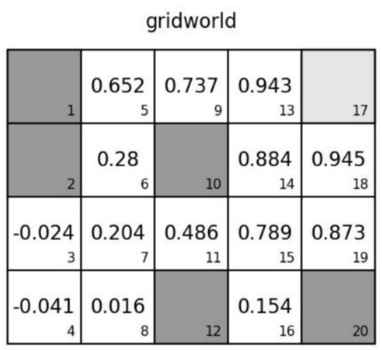
   Constant Alpha = [**0.01**,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0]

   Epsilon decay functions = [**1/((n/10)+1),** 0.9*epsilon]

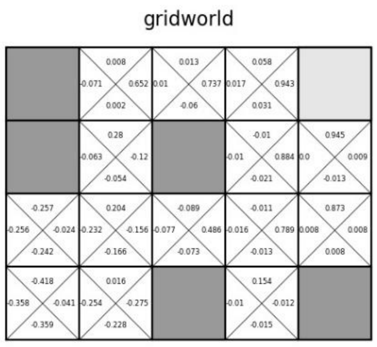
   The optimal solution (=best approximation) received by implementing the values appears in bold. After finding the optimal parameters we run the algorithm for 5000 episodes.

   The optimal solution for MC control described in figures bellow:

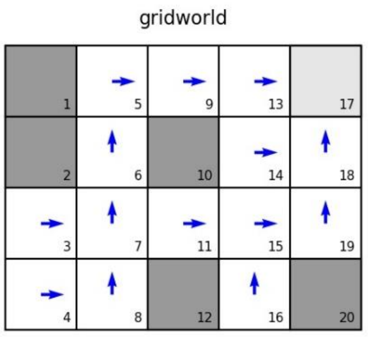
<u>Optimal state values:</u>



<u>Optimal Action state values:</u>



<u>Optimal policy:</u>



we can see that this approximation is almost converging to the optimal solution, and there for will suggest increasing number of episodes (CPU limitation)

3. <u>SARSA:</u>
   In order to find the best approximation to the optimal solution found using the DP, we tried various parameters and compared the optimal state values from SARSA to the optimal state values received from the DP. Each parameter was tested for 500 episodes.

   Tested Parameters:

   Epsilon = 0.9

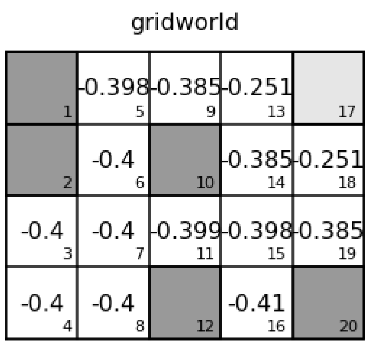
   Constant Alpha = [0.01,0.1,0.2,**0.3**,0.4,0.5,0.6,0.7,0.8,0.9,1.0]

   Epsilon decay functions = [**1/((n/10)+1)**, 0.9*epsilon]

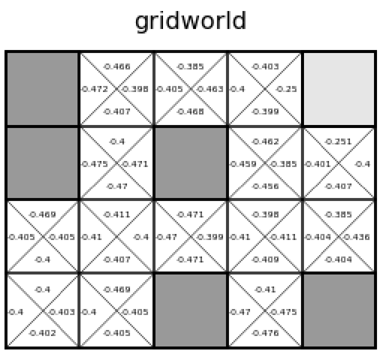
   The optimal solution (=best approximation) received by implementing the values appears in bold. After finding the optimal parameters we run the algorithm for 5000 episodes.

   The optimal solution for SARSA described in figures bellow:

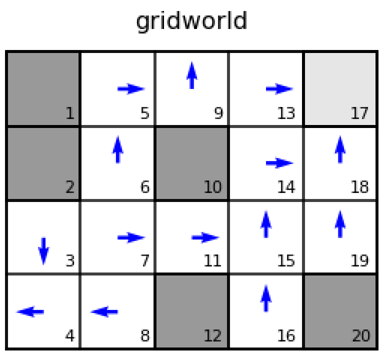
<u>Optimal state values:</u>



<u>Optimal Action state values:</u>



<u>Optimal policy:</u>



we can see that this approximation is not converging to the optimal solution, and there for will suggest increasing number of episodes (CPU limitation)

4. Q learning:

In order to find the best approximation to the optimal solution found using the DP, we tried various parameters and compared the optimal state values from Q learning to the optimal state values received from the DP. Each parameter was tested for 500 episodes.
Tested Parameters:

Epsilon = 0.9

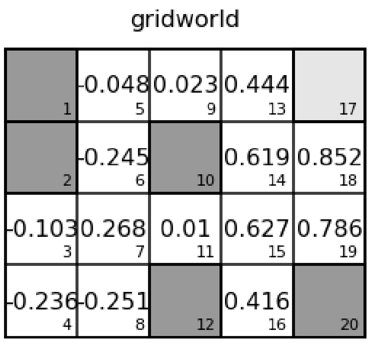
Constant Alpha = [0.01,0.1,0.2,**0.3**,0.4,0.5,0.6,0.7,0.8,0.9,1.0]

Epsilon decay functions = [**1/((n/10)+1)**, 0.9*epsilon]

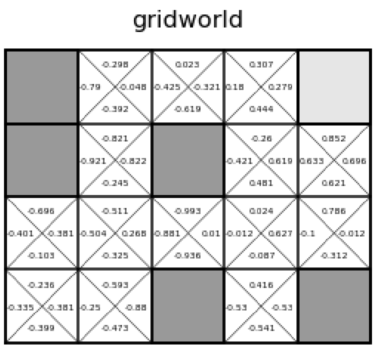
The optimal solution (=best approximation) received by implementing the values appears in bold. After finding the optimal parameters we run the algorithm for 5000 episodes.

The optimal solution for MC control described in figures bellow:

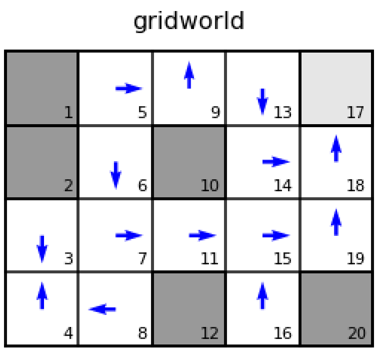
Optimal state values:



Optimal Action state values:



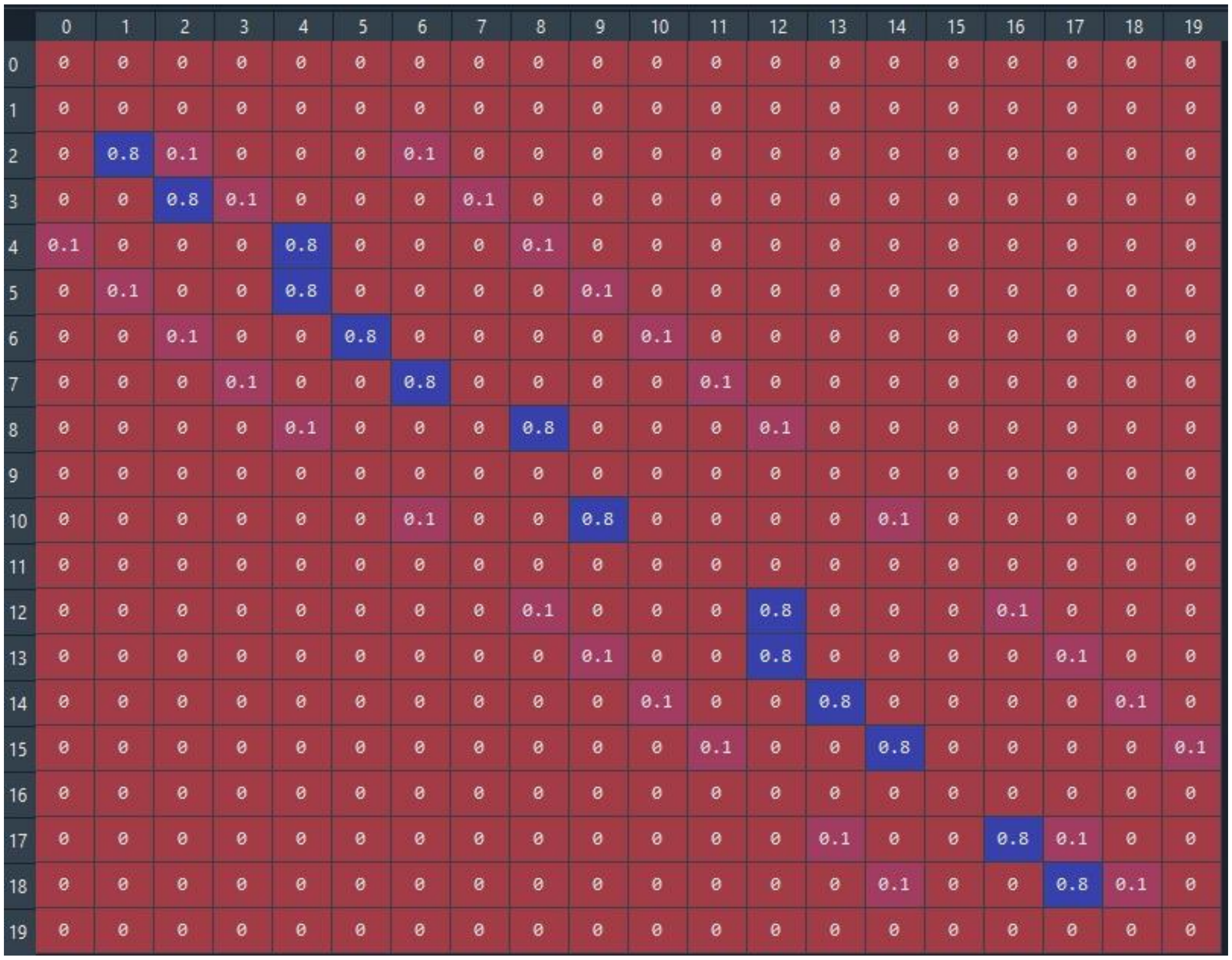
Optimal policy:



we can see that this approximation is not converging to the optimal solution, and there for will suggest increasing number of episodes (CPU limitation)

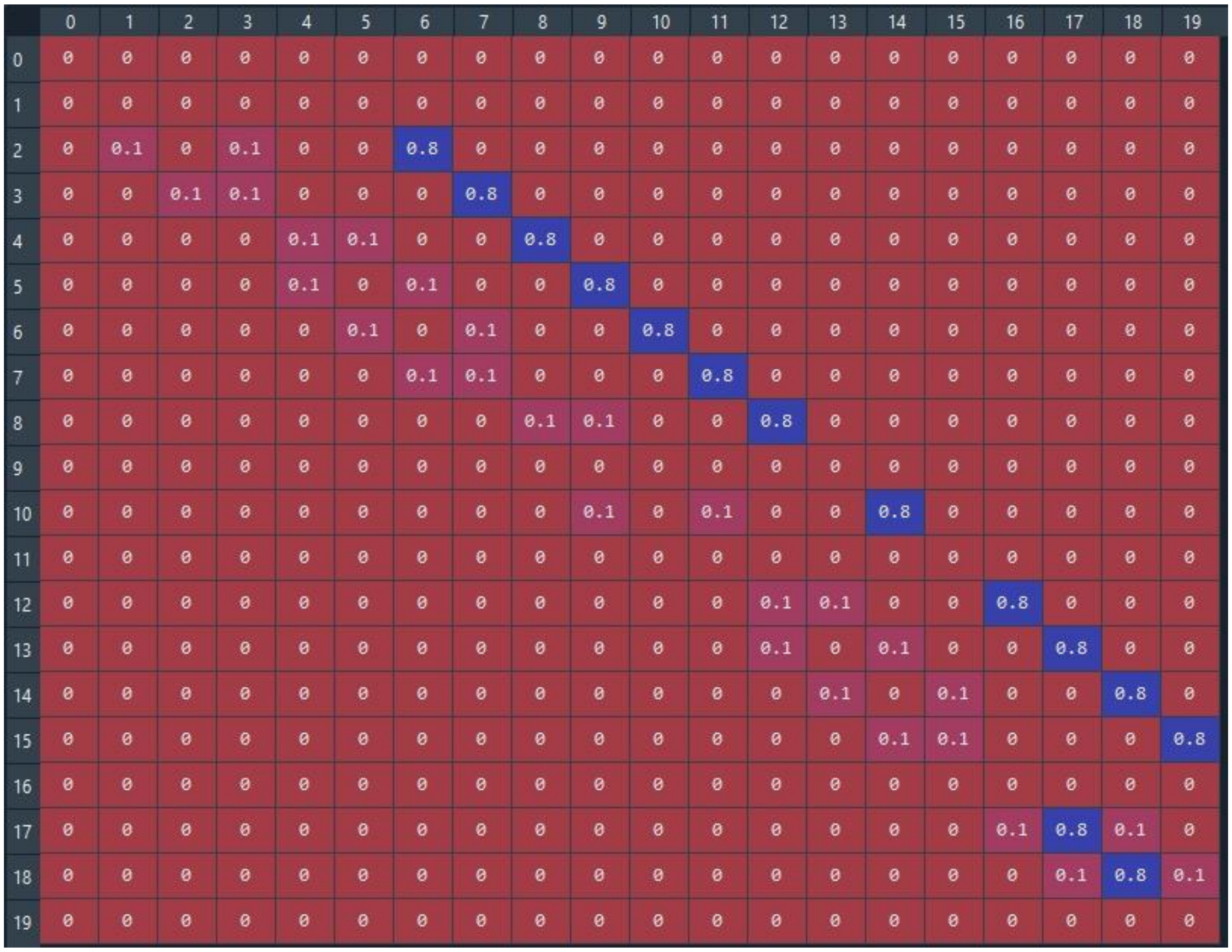
Appendix 1:

MDP transition model for Action = North: (states notation from 0 to 19)



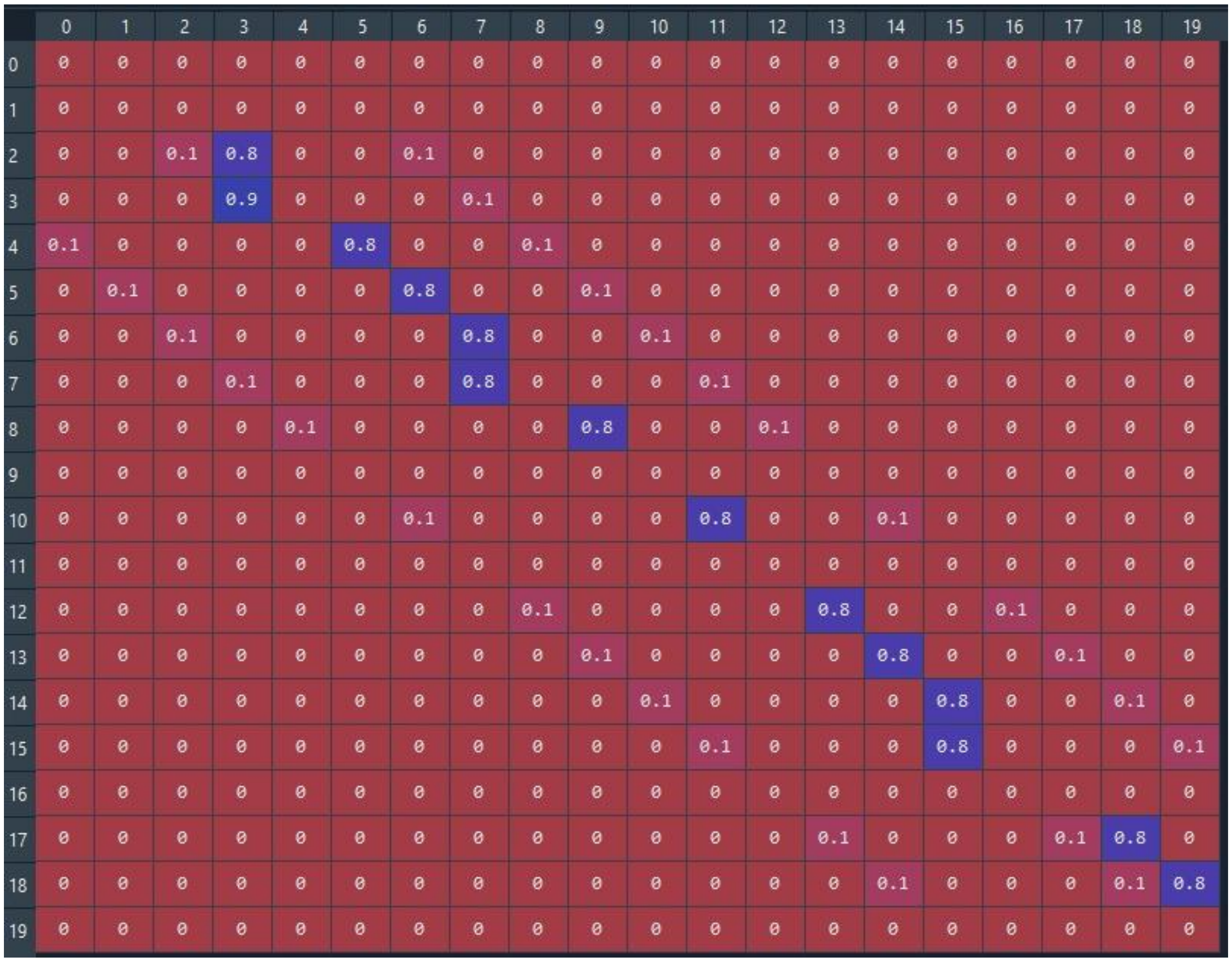
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0.8	0.1	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0.8	0.1	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0
4	0.1	0	0	0	0.8	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0
5	0	0.1	0	0	0.8	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0
6	0	0	0.1	0	0	0.8	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0
7	0	0	0	0.1	0	0	0.8	0	0	0	0	0.1	0	0	0	0	0	0	0	0
8	0	0	0	0	0.1	0	0	0	0.8	0	0	0	0.1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0.1	0	0	0.8	0	0	0	0	0.1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0.1	0	0	0	0.8	0	0	0	0.1	0	0	0
13	0	0	0	0	0	0	0	0	0	0.1	0	0	0.8	0	0	0	0	0.1	0	0
14	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0.8	0	0	0	0	0.1	0
15	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0.8	0	0	0	0	0.1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0.8	0.1	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0.8	0.1	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

MDP transition model for Action = East:



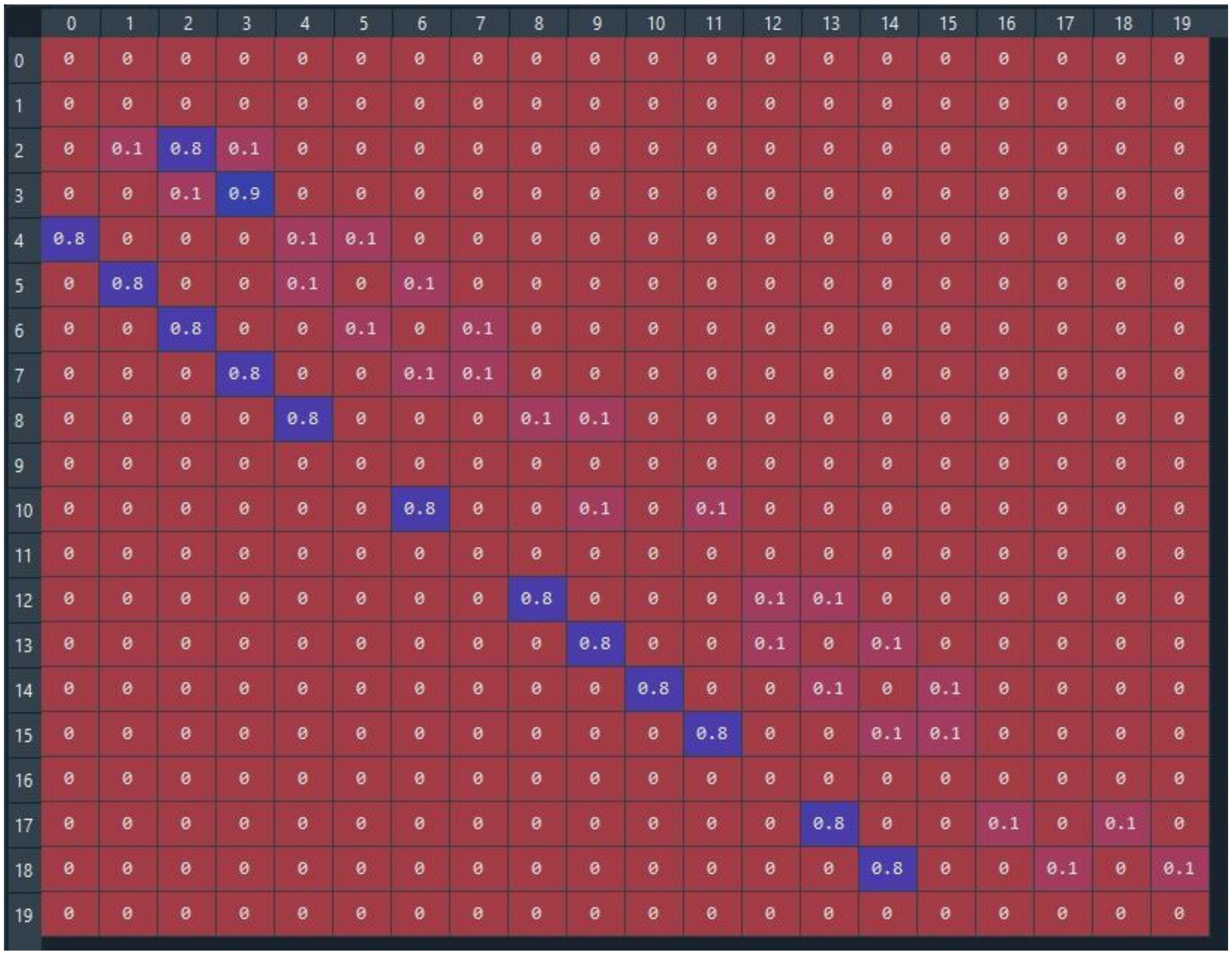
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0.1	0	0.1	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0.1	0.1	0	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0.1	0.1	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0.1	0	0.1	0	0	0.8	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0.1	0	0.1	0	0	0.8	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0.1	0.1	0	0	0	0.8	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0.1	0.1	0	0	0.8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0	0	0.8	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0	0	0.8	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0	0	0.8	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0	0	0.8	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0	0	0	0.8
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.8	0.1	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.8	0.1
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

MDP transition model for Action = South:



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0.1	0.8	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0.9	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0
4	0.1	0	0	0	0	0.8	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0
5	0	0.1	0	0	0	0	0.8	0	0	0.1	0	0	0	0	0	0	0	0	0	0
6	0	0	0.1	0	0	0	0	0.8	0	0	0.1	0	0	0	0	0	0	0	0	0
7	0	0	0	0.1	0	0	0	0.8	0	0	0	0.1	0	0	0	0	0	0	0	0
8	0	0	0	0	0.1	0	0	0	0	0.8	0	0	0.1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0.1	0	0	0	0	0.8	0	0	0.1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0.8	0	0	0.1	0	0	0
13	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0.8	0	0	0.1	0	0
14	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0.8	0	0	0.1	0
15	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.8	0	0	0	0.1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0.8	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0.8
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

MDP transition model for Action = West:



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0.1	0.8	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0.1	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0.8	0	0	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0.8	0	0	0.1	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0.8	0	0	0.1	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0.8	0	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0.8	0	0	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0.8	0	0	0.1	0	0.1	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0.8	0	0	0	0.1	0.1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0.8	0	0	0.1	0	0.1	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0.1	0	0.1	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0.1	0.1	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0.1	0	0.1	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0.1	0	0.1
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0