Report

CS6700 Spring 2024

Assignment 3
Prof Balaram Ravindran

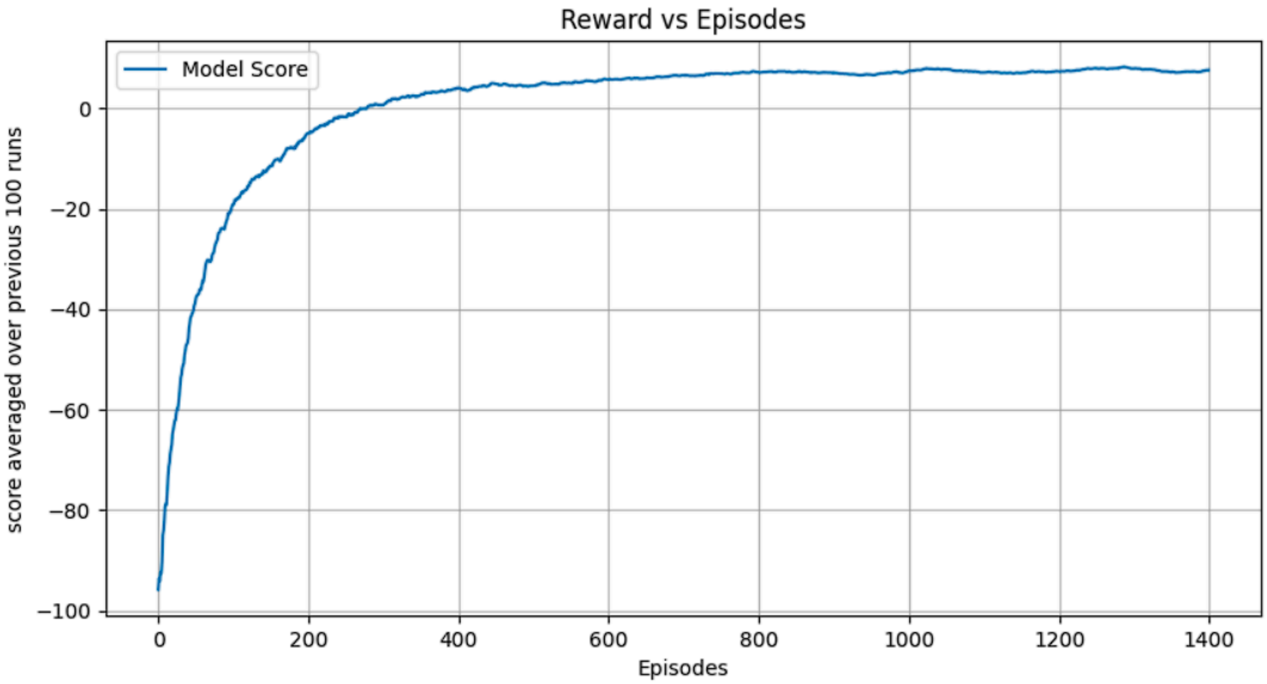
Submitted by:
Akranth Reddy me20b100@smail.iitm.ac.in
Hemanth me20b045@smail.iitm.ac.in

Date: 20-04-2024
Github Link

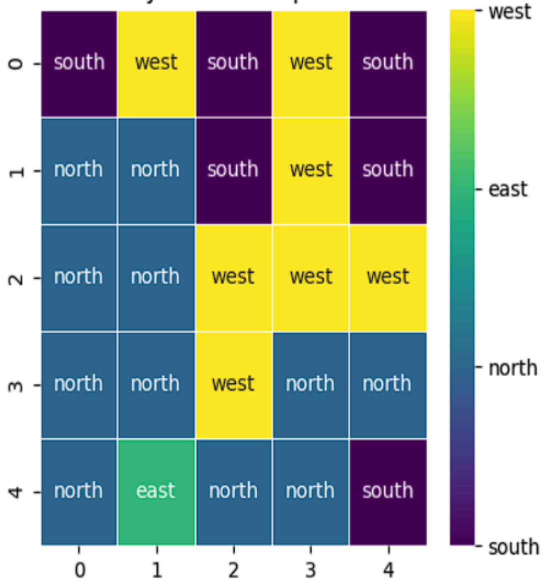
1.

SMDP Q-Learning

Reward Curves

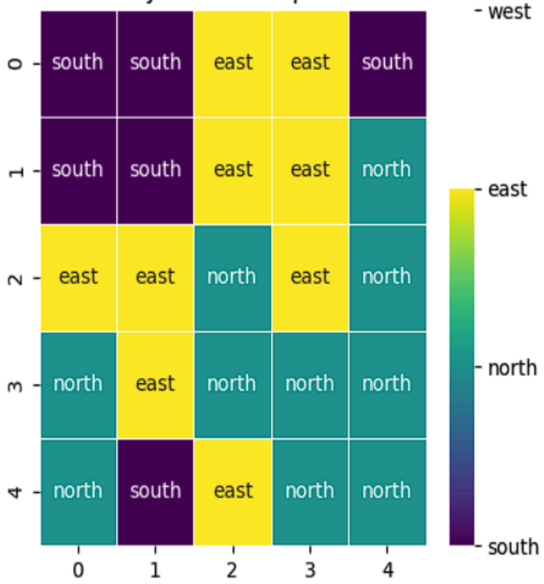


Policy learnt for option 0



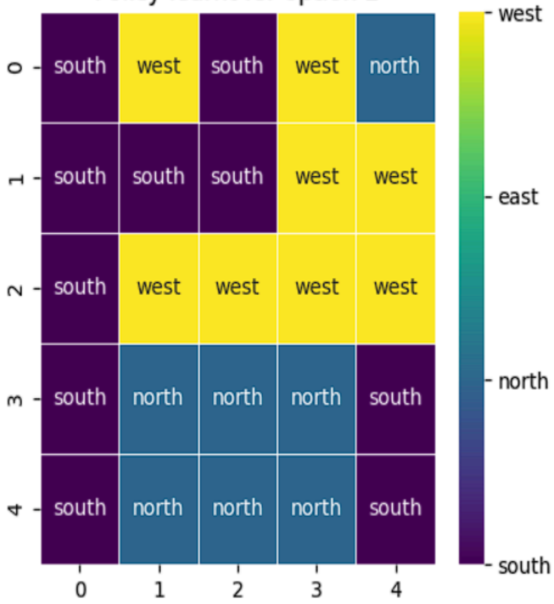
	0	1	2	3	4
0	south	west	south	west	south
1	north	north	south	west	south
2	north	north	west	west	west
3	north	north	west	north	north
4	north	east	north	north	south

Policy learnt for option 1



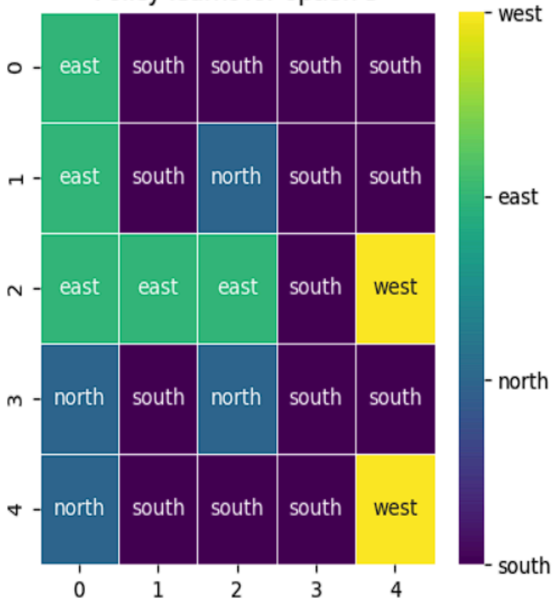
	0	1	2	3	4
0	south	south	east	east	south
1	south	south	east	east	north
2	east	east	north	east	north
3	north	east	north	north	north
4	north	south	east	north	north

Policy learnt for option 2



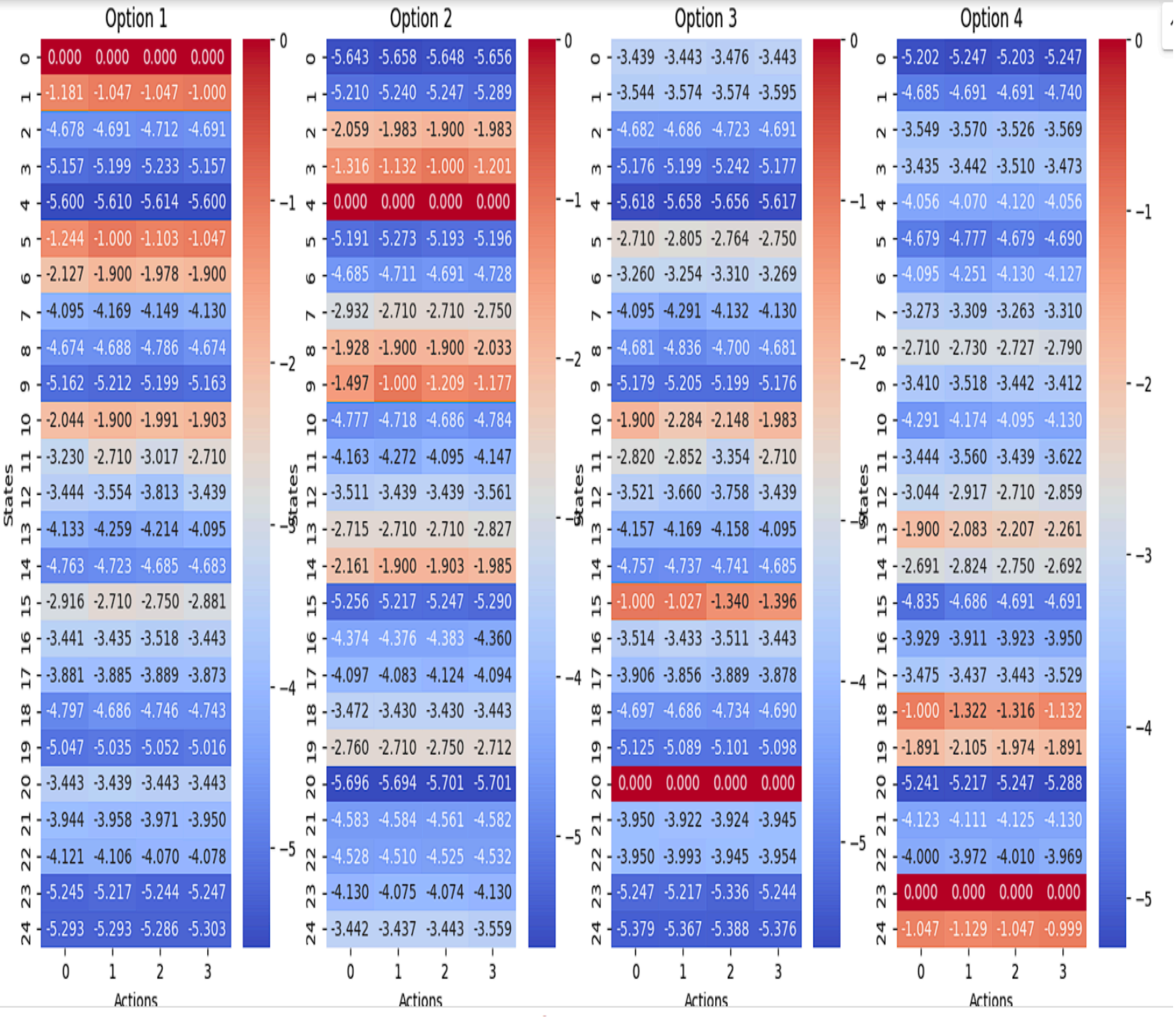
	0	1	2	3	4
0	south	west	south	west	north
1	south	south	south	west	west
2	south	west	west	west	west
3	south	north	north	north	south
4	south	north	north	north	south

Policy learnt for option 3



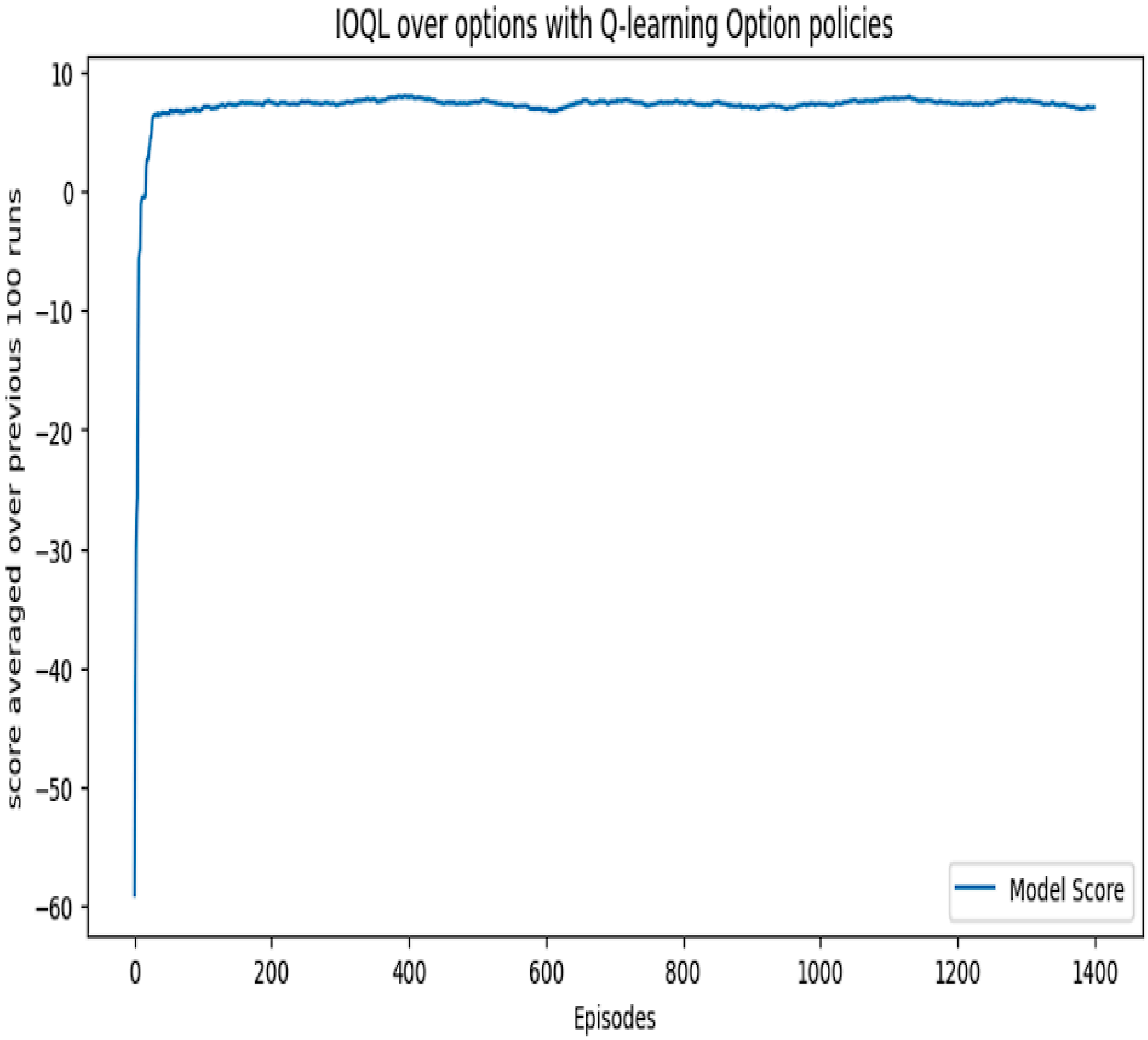
	0	1	2	3	4
0	east	south	south	south	south
1	east	south	north	south	south
2	east	east	east	south	west
3	north	south	north	south	south
4	north	south	south	south	west

Q-Values:

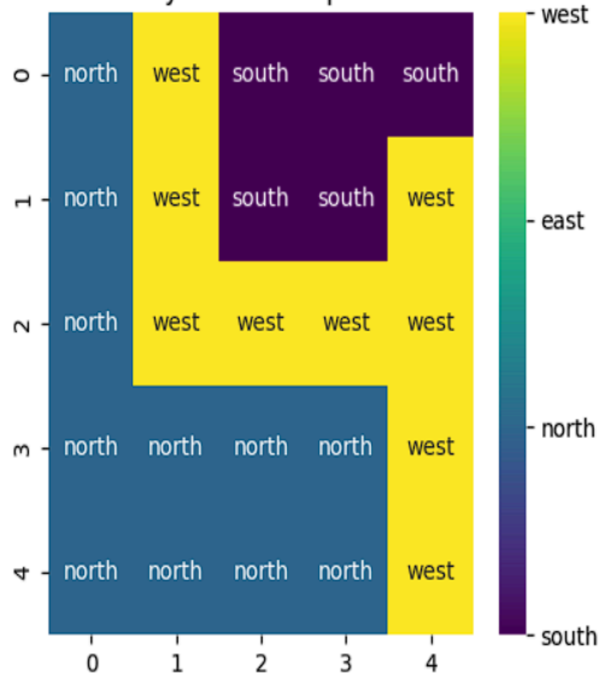

Option 1
0 0.000 0.000 0.000 0.000
1 -1.181 -1.047 -1.047 -1.000
2 -4.678 -4.691 -4.712 -4.691
3 -5.157 -5.199 -5.233 -5.157
4 -5.600 -5.610 -5.614 -5.600
5 -1.244 -1.000 -1.103 -1.047
6 -2.127 -1.900 -1.978 -1.900
7 -4.095 -4.169 -4.149 -4.130
8 -4.674 -4.688 -4.786 -4.674
9 -5.162 -5.212 -5.199 -5.163
10 -2.044 -1.900 -1.991 -1.903
11 -3.230 -2.710 -3.017 -2.710
12 -3.444 -3.554 -3.813 -3.439
13 -4.133 -4.259 -4.214 -4.095
14 -4.763 -4.723 -4.685 -4.683
15 -2.916 -2.710 -2.750 -2.881
16 -3.441 -3.435 -3.518 -3.443
17 -3.881 -3.885 -3.889 -3.873
18 -4.797 -4.686 -4.746 -4.743
19 -5.047 -5.035 -5.052 -5.016
20 -3.443 -3.439 -3.443 -3.443
21 -3.944 -3.958 -3.971 -3.950
22 -4.121 -4.106 -4.070 -4.078
23 -5.245 -5.217 -5.244 -5.247
24 -5.293 -5.293 -5.286 -5.303
States
Actions
0 1 2 3
0 -1 -2 -3 -4 -5
Option 2
0 -5.643 -5.658 -5.648 -5.656
1 -5.210 -5.240 -5.247 -5.289
2 -2.059 -1.983 -1.900 -1.983
3 -1.316 -1.132 -1.000 -1.201
4 0.000 0.000 0.000 0.000
5 -5.191 -5.273 -5.193 -5.196
6 -4.685 -4.711 -4.691 -4.728
7 -2.932 -2.710 -2.710 -2.750
8 -1.928 -1.900 -1.900 -2.033
9 -1.497 -1.000 -1.209 -1.177
10 -4.777 -4.718 -4.686 -4.784
11 -4.163 -4.272 -4.095 -4.147
12 -3.511 -3.439 -3.439 -3.561
13 -2.715 -2.710 -2.710 -2.827
14 -2.161 -1.900 -1.903 -1.985
15 -5.256 -5.217 -5.247 -5.290
16 -4.374 -4.376 -4.383 -4.360
17 -4.097 -4.083 -4.124 -4.094
18 -3.472 -3.430 -3.430 -3.443
19 -2.760 -2.710 -2.750 -2.712
20 -5.696 -5.694 -5.701 -5.701
21 -4.583 -4.584 -4.561 -4.582
22 -4.528 -4.510 -4.525 -4.532
23 -4.130 -4.075 -4.074 -4.130
24 -3.442 -3.437 -3.443 -3.559
States
Actions
0 1 2 3
0 -1 -2 -3 -4 -5
Option 3
0 -3.439 -3.443 -3.476 -3.443
1 -3.544 -3.574 -3.574 -3.595
2 -4.682 -4.686 -4.723 -4.691
3 -5.176 -5.199 -5.242 -5.177
4 -5.618 -5.658 -5.656 -5.617
5 -2.710 -2.805 -2.764 -2.750
6 -3.260 -3.254 -3.310 -3.269
7 -4.095 -4.291 -4.132 -4.130
8 -4.681 -4.836 -4.700 -4.681
9 -5.179 -5.205 -5.199 -5.176
10 -1.900 -2.284 -2.148 -1.983
11 -2.820 -2.852 -3.354 -2.710
12 -3.521 -3.660 -3.758 -3.439
13 -4.157 -4.169 -4.158 -4.095
14 -4.757 -4.737 -4.741 -4.685
15 -1.000 -1.027 -1.340 -1.396
16 -3.514 -3.433 -3.511 -3.443
17 -3.906 -3.856 -3.889 -3.878
18 -4.697 -4.686 -4.734 -4.690
19 -5.125 -5.089 -5.101 -5.098
20 0.000 0.000 0.000 0.000
21 -3.950 -3.922 -3.924 -3.945
22 -3.950 -3.993 -3.945 -3.954
23 -5.247 -5.217 -5.336 -5.244
24 -5.379 -5.367 -5.388 -5.376
States
Actions
0 1 2 3
0 -1 -2 -3 -4 -5
Option 4
0 -5.202 -5.247 -5.203 -5.247
1 -4.685 -4.691 -4.691 -4.740
2 -3.549 -3.570 -3.526 -3.569
3 -3.435 -3.442 -3.510 -3.473
4 -4.056 -4.070 -4.120 -4.056
5 -4.679 -4.777 -4.679 -4.690
6 -4.095 -4.251 -4.130 -4.127
7 -3.273 -3.309 -3.263 -3.310
8 -2.710 -2.730 -2.727 -2.790
9 -3.410 -3.518 -3.442 -3.412
10 -4.291 -4.174 -4.095 -4.130
11 -3.444 -3.560 -3.439 -3.622
12 -3.044 -2.917 -2.710 -2.859
13 -1.900 -2.083 -2.207 -2.261
14 -2.691 -2.824 -2.750 -2.692
15 -4.835 -4.686 -4.691 -4.691
16 -3.929 -3.911 -3.923 -3.950
17 -3.475 -3.437 -3.443 -3.529
18 -1.000 -1.322 -1.316 -1.132
19 -1.891 -2.105 -1.974 -1.891
20 -5.241 -5.217 -5.247 -5.288
21 -4.123 -4.111 -4.125 -4.130
22 -4.000 -3.972 -4.010 -3.969
23 0.000 0.000 0.000 0.000
24 -1.047 -1.129 -1.047 -0.999
States
Actions
0 1 2 3
0 -1 -2 -3 -4 -5

Intra-Option Q-Learning

Reward Curves

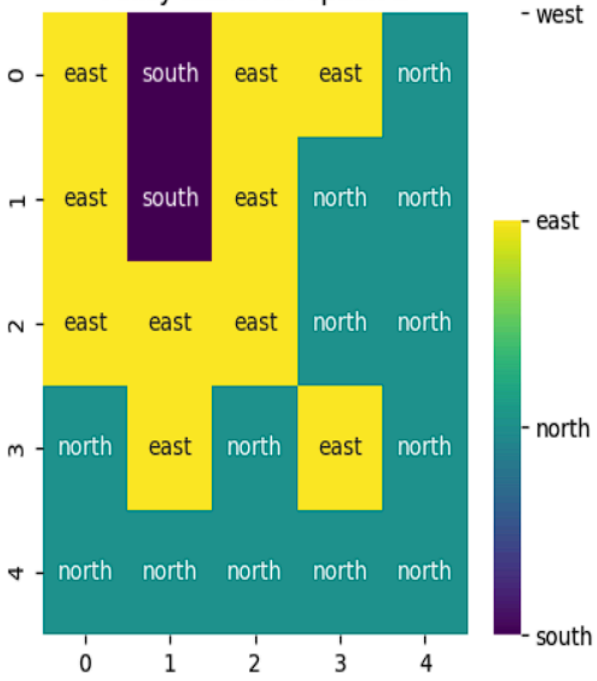


Policy learnt for option 0



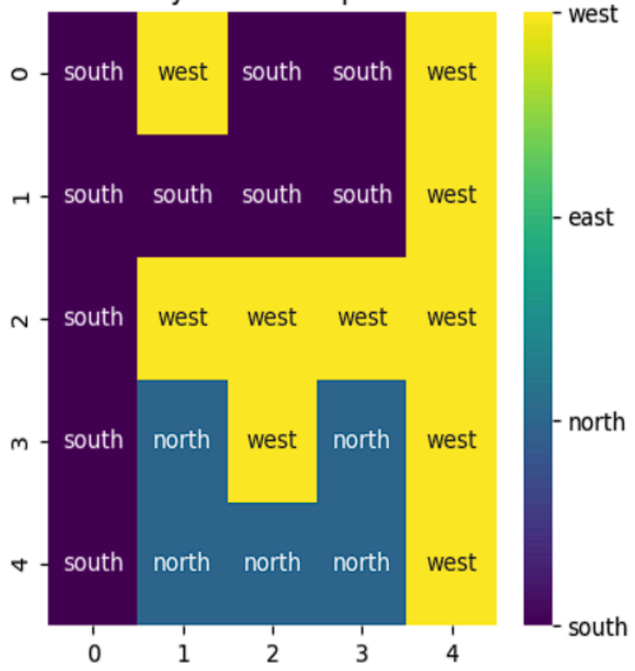
	0	1	2	3	4
0	north	west	south	south	south
1	north	west	south	south	west
2	north	west	west	west	west
3	north	north	north	north	west
4	north	north	north	north	west

Policy learnt for option 1



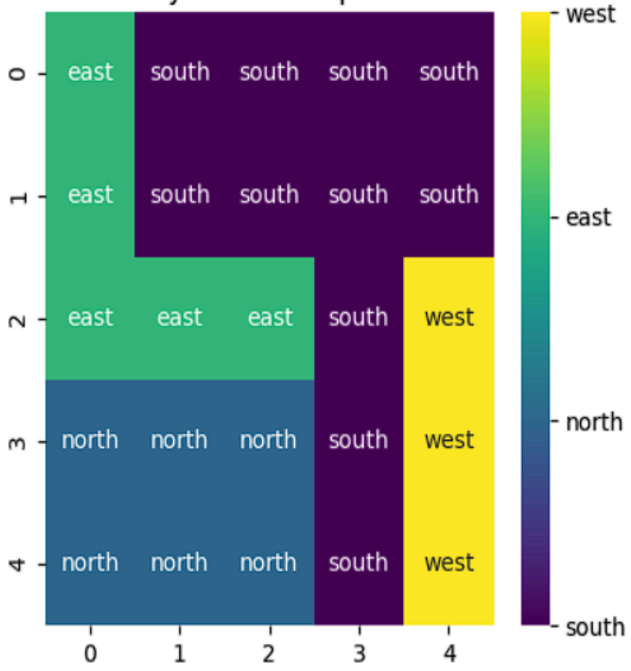
	0	1	2	3	4
0	east	south	east	east	north
1	east	south	east	north	north
2	east	east	east	north	north
3	north	east	north	east	north
4	north	north	north	north	north

Policy learnt for option 2



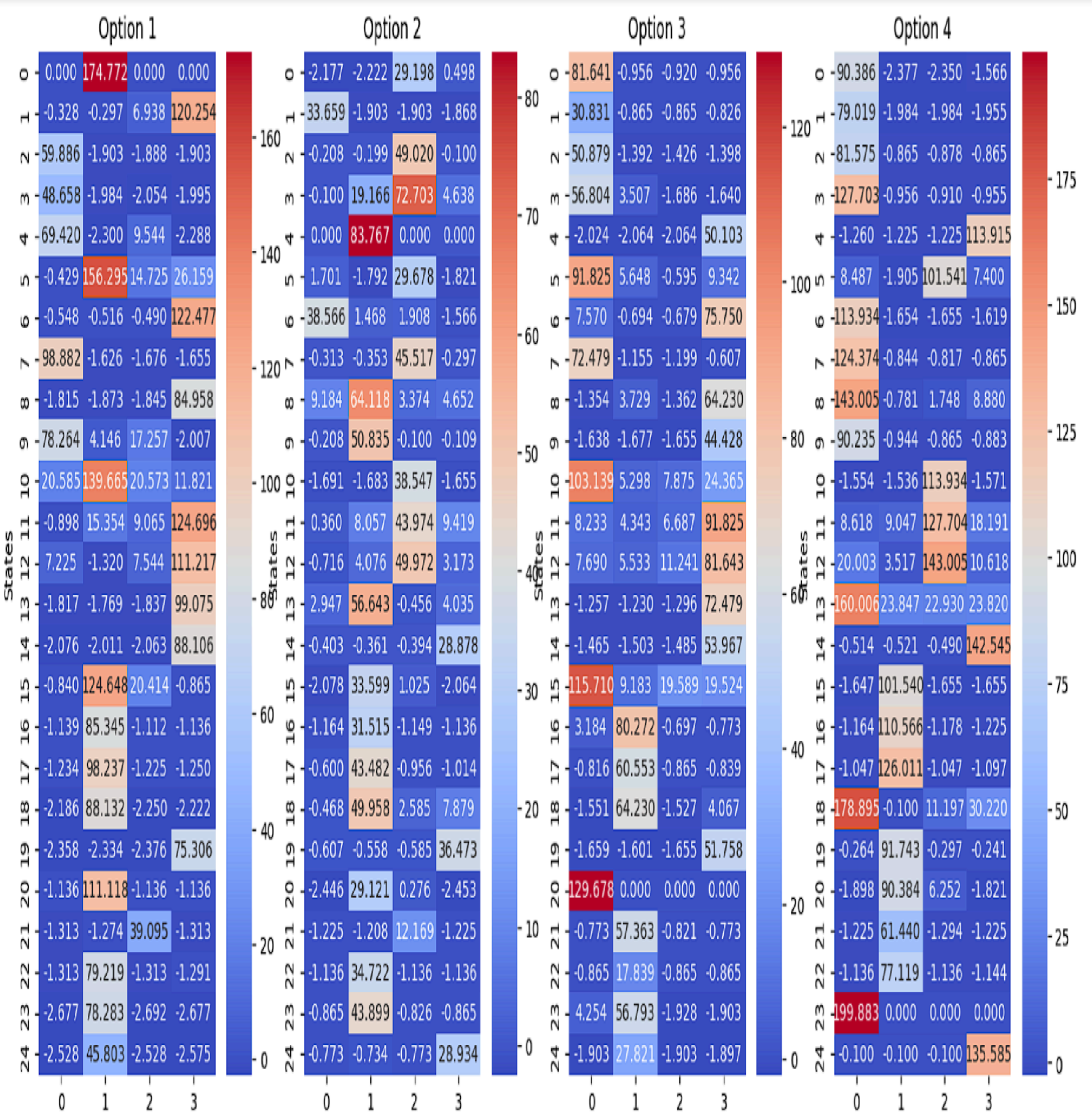
	0	1	2	3	4
0	south	west	south	south	west
1	south	south	south	south	west
2	south	west	west	west	west
3	south	north	west	north	west
4	south	north	north	north	west

Policy learnt for option 3



	0	1	2	3	4
0	east	south	south	south	south
1	east	south	south	south	south
2	east	east	east	south	west
3	north	north	north	south	west
4	north	north	north	south	west

Q-Values



Option 1

States	0	1	2	3
0	0.000	174.772	0.000	0.000
1	-0.328	-0.297	6.938	120.254
2	59.886	-1.903	-1.888	-1.903
3	48.658	-1.984	-2.054	-1.995
4	69.420	-2.300	9.544	-2.288
5	-0.429	156.295	14.725	26.159
6	-0.548	-0.516	-0.490	122.477
7	98.882	-1.626	-1.676	-1.655
8	-1.815	-1.873	-1.845	84.958
9	78.264	4.146	17.257	-2.007
10	20.585	139.665	20.573	11.821
11	-0.898	15.354	9.065	124.696
12	7.225	-1.320	7.544	111.217
13	-1.817	-1.769	-1.837	99.075
14	-2.076	-2.011	-2.063	88.106
15	-0.840	124.648	20.414	-0.865
16	-1.139	85.345	-1.112	-1.136
17	-1.234	98.237	-1.225	-1.250
18	-2.186	88.132	-2.250	-2.222
19	-2.358	-2.334	-2.376	75.306
20	-1.136	111.118	-1.136	-1.136
21	-1.313	-1.274	39.095	-1.313
22	-1.313	79.219	-1.313	-1.291
23	-2.677	78.283	-2.692	-2.677
24	-2.528	45.803	-2.528	-2.575

Option 2

States	0	1	2	3
0	-2.177	-2.222	29.198	0.498
1	33.659	-1.903	-1.903	-1.868
2	-0.208	-0.199	49.020	-0.100
3	-0.100	19.166	72.703	4.638
4	0.000	83.767	0.000	0.000
5	1.701	-1.792	29.678	-1.821
6	38.566	1.468	1.908	-1.566
7	-0.313	-0.353	45.517	-0.297
8	9.184	64.118	3.374	4.652
9	-0.208	50.835	-0.100	-0.109
10	-1.691	-1.683	38.547	-1.655
11	0.360	8.057	43.974	9.419
12	-0.716	4.076	49.972	3.173
13	2.947	56.643	-0.456	4.035
14	-0.403	-0.361	-0.394	28.878
15	-2.078	33.599	1.025	-2.064
16	-1.164	31.515	-1.149	-1.136
17	-0.600	43.482	-0.956	-1.014
18	-0.468	49.958	2.585	7.879
19	-0.607	-0.558	-0.585	36.473
20	-2.446	29.121	0.276	-2.453
21	-1.225	-1.208	12.169	-1.225
22	-1.136	34.722	-1.136	-1.136
23	-0.865	43.899	-0.826	-0.865
24	-0.773	-0.734	-0.773	28.934

Option 3

States	0	1	2	3
0	81.641	-0.956	-0.920	-0.956
1	30.831	-0.865	-0.865	-0.826
2	50.879	-1.392	-1.426	-1.398
3	56.804	3.507	-1.686	-1.640
4	-2.024	-2.064	-2.064	50.103
5	91.825	5.648	-0.595	9.342
6	7.570	-0.694	-0.679	75.750
7	72.479	-1.155	-1.199	-0.607
8	-1.354	3.729	-1.362	64.230
9	-1.638	-1.677	-1.655	44.428
10	103.139	5.298	7.875	24.365
11	8.233	4.343	6.687	91.825
12	7.690	5.533	11.241	81.643
13	-1.257	-1.230	-1.296	72.479
14	-1.465	-1.503	-1.485	53.967
15	115.710	9.183	19.589	19.524
16	3.184	80.272	-0.697	-0.773
17	-0.816	60.553	-0.865	-0.839
18	-1.551	64.230	-1.527	4.067
19	-1.659	-1.601	-1.655	51.758
20	129.678	0.000	0.000	0.000
21	-0.773	57.363	-0.821	-0.773
22	-0.865	17.839	-0.865	-0.865
23	4.254	56.793	-1.928	-1.903
24	-1.903	27.821	-1.903	-1.897

Option 4

States	0	1	2	3
0	90.386	-2.377	-2.350	-1.566
1	79.019	-1.984	-1.984	-1.955
2	81.575	-0.865	-0.878	-0.865
3	127.703	-0.956	-0.910	-0.955
4	-1.260	-1.225	-1.225	113.915
5	8.487	-1.905	101.541	7.400
6	113.934	-1.654	-1.655	-1.619
7	124.374	-0.844	-0.817	-0.865
8	143.005	-0.781	1.748	8.880
9	90.235	-0.944	-0.865	-0.883
10	-1.554	-1.536	113.934	-1.571
11	8.618	9.047	127.704	18.191
12	20.003	3.517	143.005	10.618
13	160.006	23.847	22.930	23.820
14	-0.514	-0.521	-0.490	142.545
15	-1.647	101.540	-1.655	-1.655
16	-1.164	110.566	-1.178	-1.225
17	-1.047	126.011	-1.047	-1.097
18	178.895	-0.100	11.197	30.220
19	-0.264	91.743	-0.297	-0.241
20	-1.898	90.384	6.252	-1.821
21	-1.225	61.440	-1.294	-1.225
22	-1.136	77.119	-1.136	-1.144
23	199.883	0.000	0.000	0.000
24	-0.100	-0.100	-0.100	135.585

2.

SMDP:

- The policy learnt for option 0 is to move east if the passenger is at location 1 or 2, and to move north if the passenger is at location 3. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move south to drop them off at location 0.

- The policy learnt for option 1 is to move west if the passenger is at location 0 or 3, and to move south if the passenger is at location 2. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move north to drop them off at location 1.

- The policy learnt for option 2 is to move north if the passenger is at location 0 or 1, and to move east if the passenger is at location 3. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move west to drop them off at location 2.

- The policy learnt for option 3 is to move south if the passenger is at location 0 or 2, and to move west if the passenger is at location 1. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move east to drop them off at location 3.

Intra-Option Q-Learning:

- The policy learnt for option 0 is to move east if the passenger is at location 1 or 2, and to move north if the passenger is at location 3. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move south to drop them off at location 0.

- The policy learnt for option 1 is to move west if the passenger is at location 0 or 3, and to move south if the passenger is at location 2. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move north to drop them off at location 1.

- The policy learnt for option 2 is to move north if the passenger is at location 0 or 1, and to move east if the passenger is at location 3. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move west to drop them off at location 2.

- The policy learnt for option 3 is to move south if the passenger is at location 0 or 2, and to move west if the passenger is at location 1. This is because the taxi needs to move towards the passenger to pick them up.
- Once the passenger is picked up, the policy is to move east to drop them off at location **3.**

Reasoning:

- Both the SMDP and intra-option Q-learning algorithms learn the optimal policy for the Taxi environment.
- This is because both algorithms take into account the reward structure of the environment and learn to maximize the expected return.
- The main difference between the two algorithms is that the intra-option Q-learning algorithm can learn about the values of executing certain options without ever executing those options.
- This is because the intra-option Q-learning algorithm uses a special temporal-difference method to learn about the consequences of one policy while actually behaving according to another, potentially different policy.
- This allows the intra-option Q-learning algorithm to learn more efficiently than the SMDP algorithm.
- The SMDP algorithm learns the policy for each option by iteratively updating the Q-values for each state-action pair. The Q-values represent the expected return for taking a particular action in a given state. The algorithm then uses the Q-values to select the action with the highest expected return in each state.
- The Intra-Option Q-Learning algorithm also learns the policy for each option by iteratively updating the Q-values for each state-action pair. However, the Intra-Option Q-Learning algorithm also uses the Q-values to learn about the value of executing certain options without ever executing those options. This is because the Intra-Option Q-Learning algorithm can learn from the experience of other options.
- Both the SMDP and Intra-Option Q-Learning algorithms learn the optimal policy for the task of picking up and dropping off passengers in the Taxi environment. This is because both algorithms take into account the reward structure of the environment and learn to maximize the expected return.

3.

Comparison between SMDP Q-Learning and intra-option Q-Learning algorithms:

SMDP Q-Learning:

- Learns the policy for each option by iteratively updating the Q-values for each state-action pair.
- Uses the Q-values to select the action with the highest expected return in each state.

Intra-option Q-Learning:

- Also learns the policy for each option by iteratively updating the Q-values for each state-action pair.
- Uses the Q-values to learn about the value of executing certain options without ever executing those options.
- Can learn from the experience of other options.

Improvement with intra-option Q-Learning:

- Yes, intra-option Q-Learning can improve performance over SMDP Q-Learning.
- This is because intra-option Q-Learning can learn from the experience of other options, which can help it to find a better policy.
- Intra-option Q-Learning can also learn about the value of executing certain options without ever executing those options, which can help it to make better decisions about which option to execute.

Briefly:

- SMDP Q-Learning learns the policy for each option independently.
- Intra-option Q-Learning can learn from the experience of other options.
- Intra-option Q-Learning can improve performance over SMDP Q-Learning.