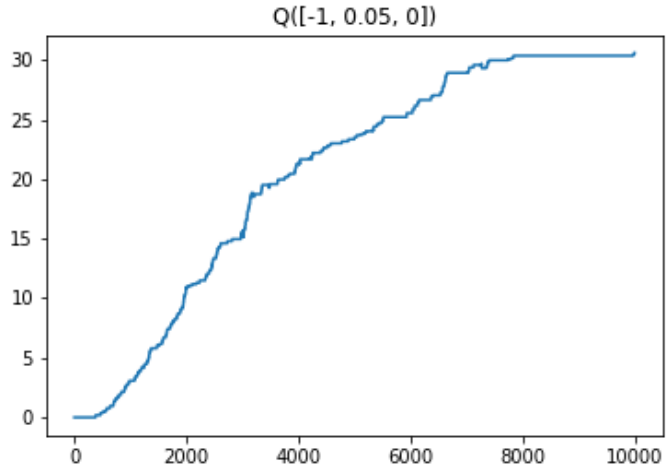
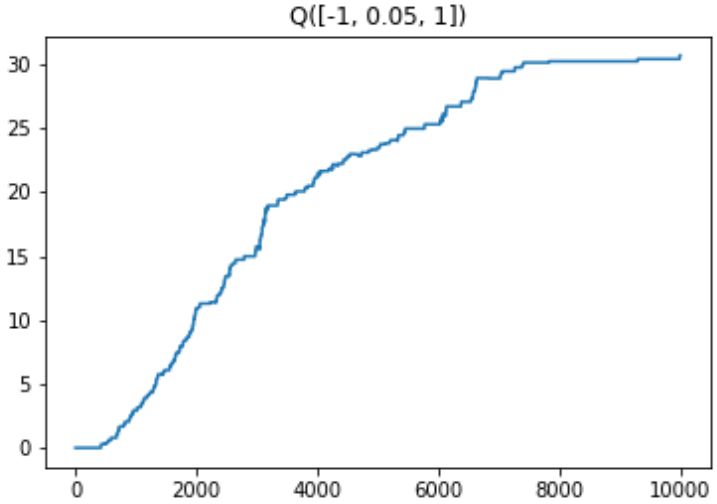
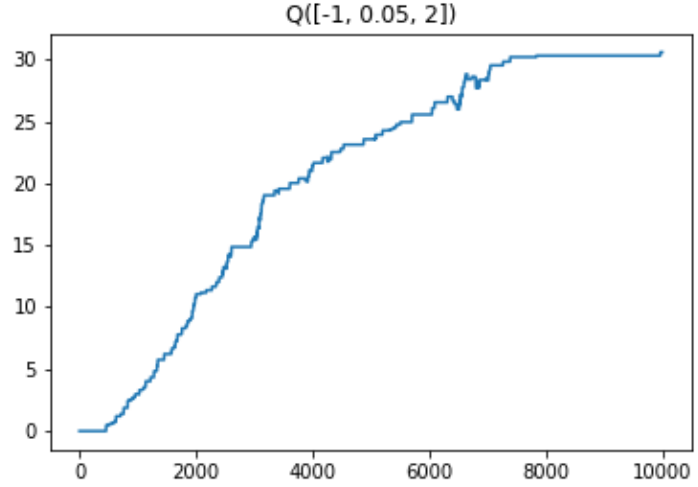
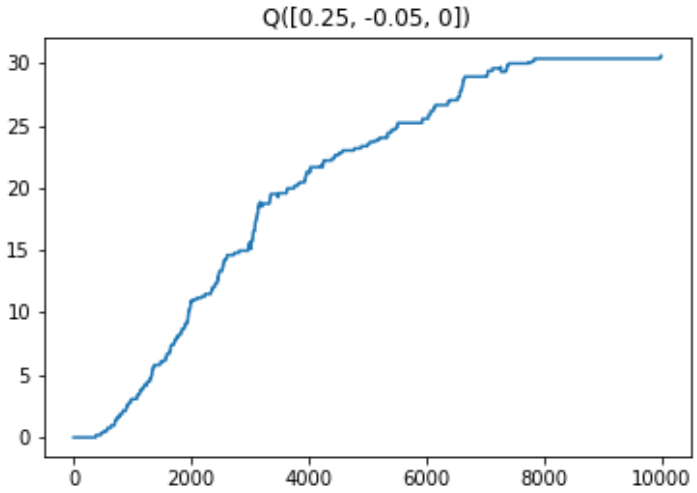
State 2:

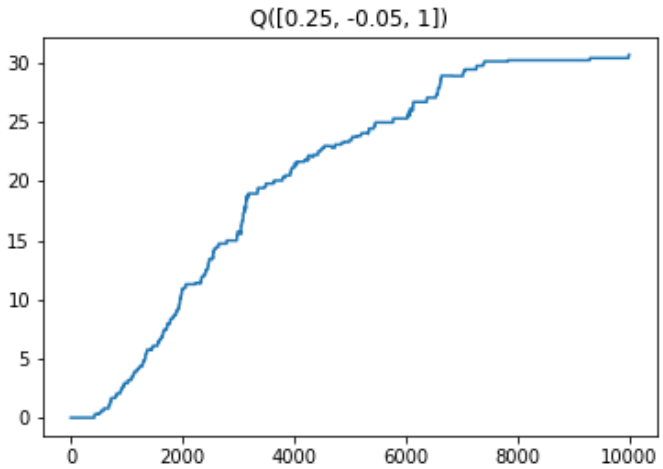

Q([-1, 0.05, 0])
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000

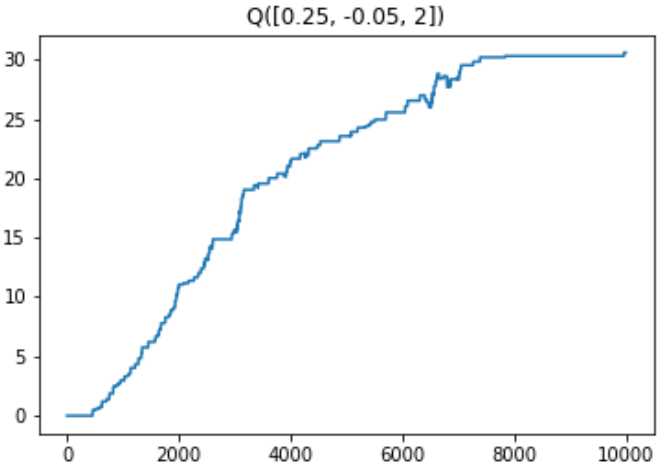

Q([-1, 0.05, 1])
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000


Q([-1, 0.05, 2])
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000

State 3:

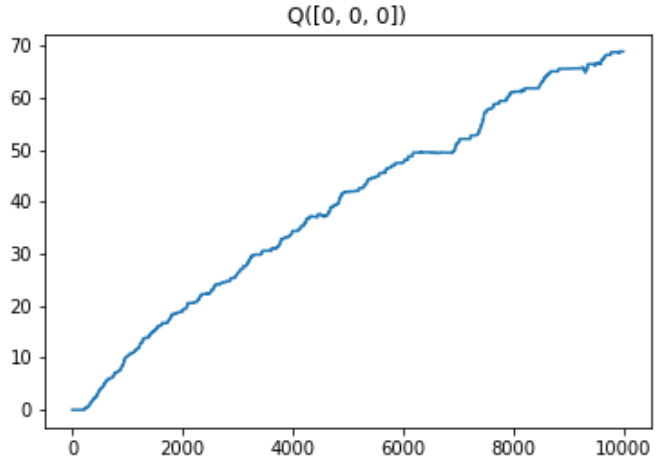

Q([0.25, -0.05, 0])
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000

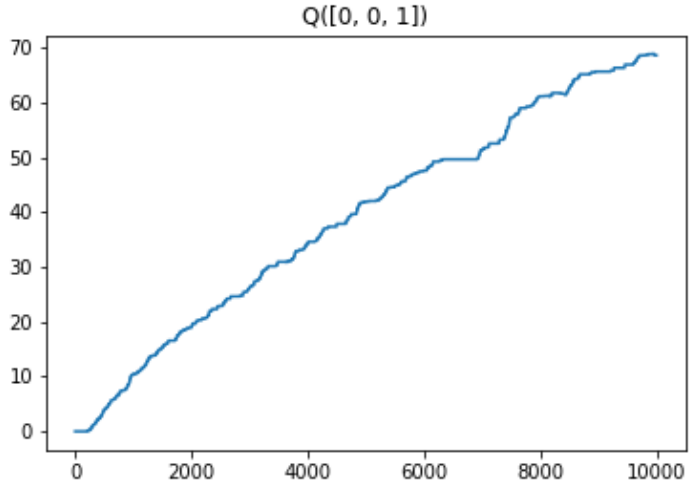

Q([0.25, -0.05, 1])
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000

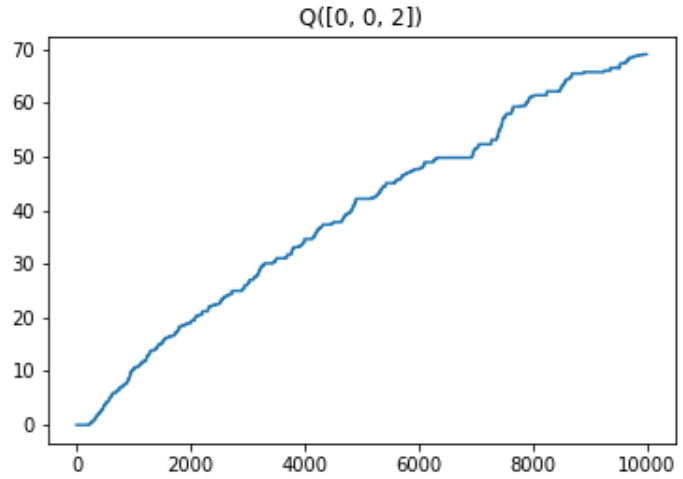

Q([0.25, -0.05, 2])
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000

Off-policy TD with Q learning:

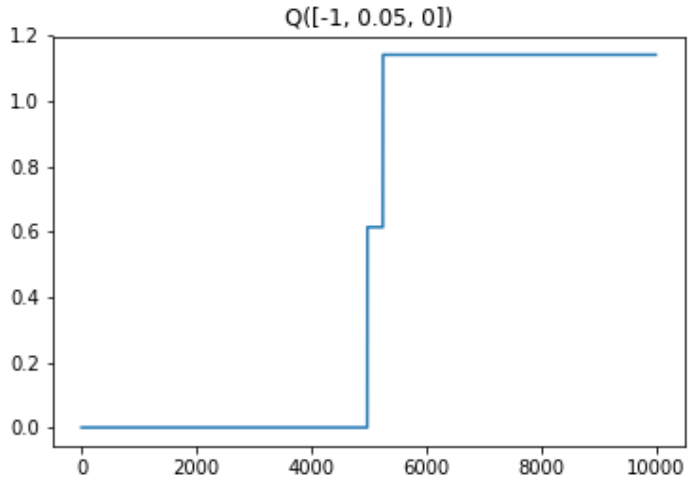
State 1:

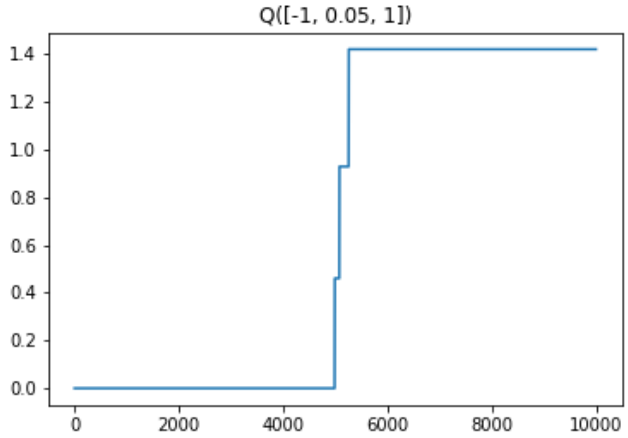


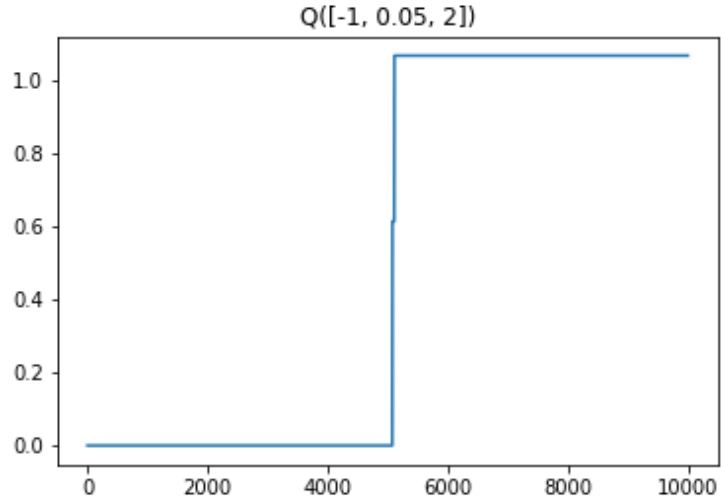




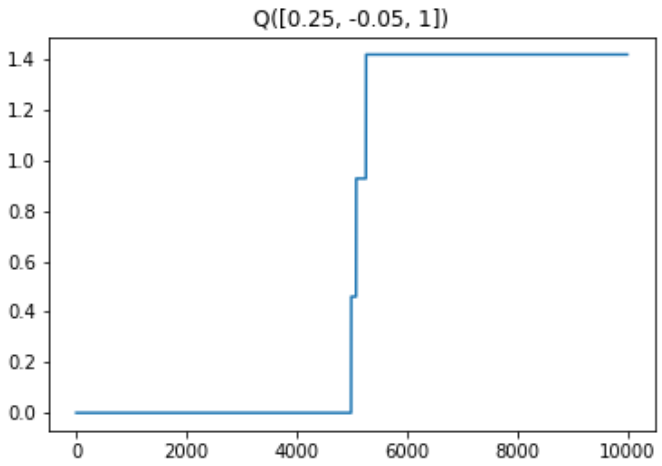
State 2:

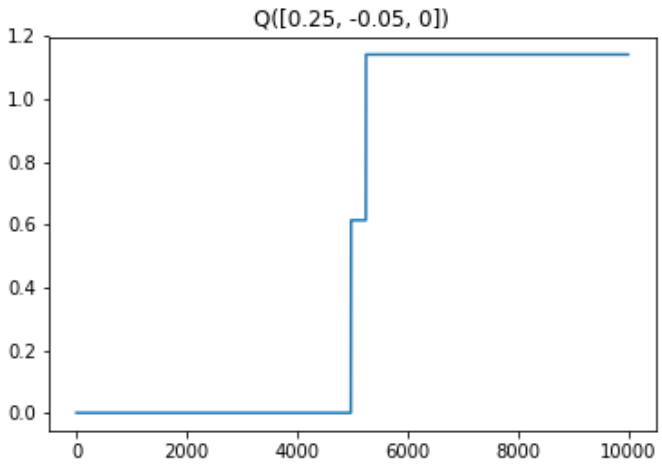

Q([-1, 0.05, 0])
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000

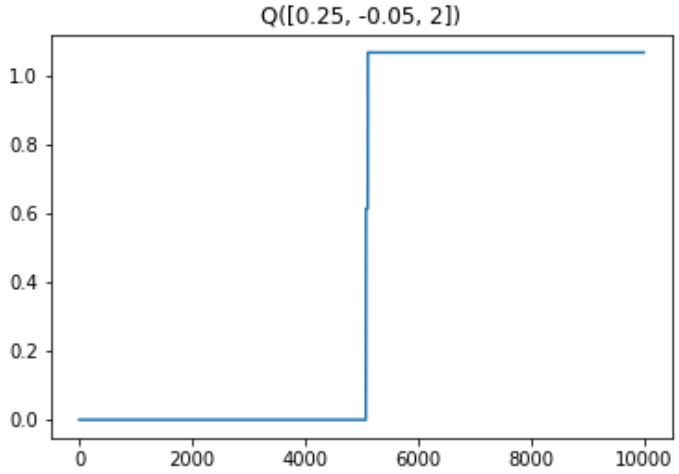

Q([-1, 0.05, 1])
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000


Q([-1, 0.05, 2])
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000

State 3:


Q([0.25, -0.05, 1])
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000


Q([0.25, -0.05, 0])
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000


Q([0.25, -0.05, 2])
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000

Optimized Policy: (x-axis is velocity [from -0.7 to 0.7], y-axis is position [from -1.2 to 0.6])