Assignment 2

Yinan Shi
Matrikelnummer: 03692624

Exercise 1

Learned GMM parameters:

1) Prior

0.2400	0.2011	0.2972	0.2617

2) Means

X1	-0.0432	-0.0147	-0.0194	0.0262
X2	0.0446	-0.0796	-0.0166	0.0617

3) Covariance matrices

0.00017479	0.00026154
0.00026154	0.00039754

0.00039440	0.00021664
0.00021664	0.00012757

0.00074372	-0.00059168
-0.00059168	0.00061026

0.0011	-0.00042436
-0.00042436	0.00024312

Exercise 2

All the sequences are classified to gesture 2.

Exercise 3

Task 2

1) Reward Matrix

0	-1	0	-1
0	0	-1	-1
0	0	-1	-1
0	-1	0	-1
-1	-1	0	0
0	-1	0	-1
0	-1	0	-1
-1	1	0	0
-1	-1	0	0
0	-1	0	-1
0	-1	0	-1
-1	1	0	0
0	-1	0	-1
0	0	-1	1
0	0	-1	1
0	1	0	1

2)

In my function, I use a γ with a value of 0.9. For a smaller γ the iteration will be faster to converge. Because the Value Function V can be written as the expected total future reward:

$$V^{\pi}(s) = E\left[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots \mid s_0 = s, \pi\right]$$

So the discount factor controls to what extend further rewards will affect the value function. Reducing the discount factor would reduce the affection of future rewards and care more about the current rewards. This would lead to a faster convergence, but the learned policy might not be the best. This is the theoretical implementation.

In my function, I have tried five different values of γ, which is 0.9, 0.7, 0.5, 0.3, 0.1.

The outcomes are all the same, and the iteration steps have only a slight difference.

3)

Approximately, it only needs 6 iterations to converge. Sometimes it will be a little bit longer, with 7 steps, while sometimes also a little bit faster, with 4 steps.

The result of *WalkPolicyIteration(10)*:

The result of *WalkPolicyIteration(3)*:

Task 3

4)

Here, I use a ε with value 0.2, and a α with value 0.8.

5)

If a pure greedy policy is used, the outcome will never be correct no matter how many iteration steps you set. Because the Q-Learning will only converge to a local optimal policy. It has no chance (ε) to explore a more global optimal policy.

The value of ε matters. If the value of ε increases, it will also increase the probability of random choice, which makes it more easier to explore the better optimal policy, and increases the uncertainty as well. If the value of ε decreases, it has a low ability of discovering better, more global optimal policies, which means it needs more steps for a good outcome.

6)

With the value I have shown in 4), it needs approximately 1500 steps to have a good outcome. More steps will lead to the same result.

7)

The result of *WalkQLearning(5)*:

The result of *WalkQLearning(12)*: