Human-centered Assistive Robotics Technische Universität München Prof. Dr.-Ing. Dongheui Lee	**MACHINE LEARNING IN ROBOTICS** Assignment 2 Tianming Qiu 03686061

Exercise 1

GMM Parameters	Prior	Mean	Covariance Matrix
Cluster 1	0.2407	$(-0.0432 \quad 0.0446)^T$	$\begin{pmatrix} 0.1780e-03 & 0.2664e-03 \\ 0.2664e-03 & 0.4051e-03 \end{pmatrix}$
Cluster 2	0.2016	$(-0.0146 \quad -0.0796)^T$	$\begin{pmatrix} 0.4026e-03 & 0.2215e-03 \\ 0.2215e-03 & 0.1307e-03 \end{pmatrix}$
Cluster 3	0.2612	$(0.0263 \quad 0.0617)^T$	$\begin{pmatrix} 0.0011 & -0.0004 \\ -0.0004 & 0.0002 \end{pmatrix}$
Cluster 4	0.2964	$(-0.0194 \quad -0.0166)^T$	$\begin{pmatrix} 0.7505e-03 & -0.5964e-03 \\ -0.5964e-03 & 0.6148e-03 \end{pmatrix}$

Exercise 2

The log-likelihood results are:

[-511.407, -570.67, -387.917, -427.307, -437.599, -426.178, -473.303, -400.288, -377.18, -401.06]

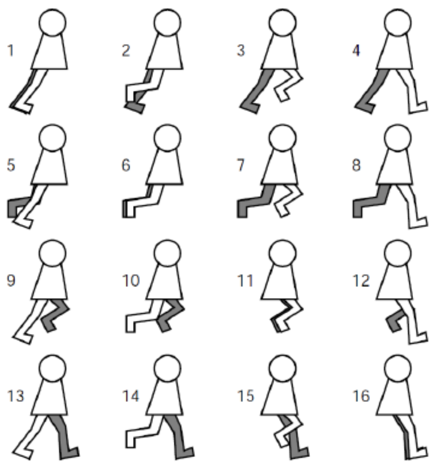
So all of them are classified as gesture 2.

Exercise 3

Policy Iteration

1. Reward matrix:

State	R u/d	R f/b	L u/d	L f/b
1	-1	1	-1	1
2	-1	1	-1	-1
3	1	-1	-1	-1
4	-1	-1	1	-1
5	-1	-1	-1	1
6	1	-1	1	-1
7	1	-1	-1	-1
8	-1	1	-1	-1
9	-1	-1	1	-1
10	-1	-1	1	-1
11	1	-1	1	-1
12	-1	1	-1	-1
13	1	-1	-1	-1
14	-1	-1	-1	1
15	-1	-1	-1	1
16	-1	1	-1	1



2. $\gamma = 0.8$ is adopted. γ represents the influence of the future reward. When γ increases/decreases, more/less influence of future reward will be considered. In this problem the result of changing is not obvious. There are two reasons: one is the iteration times are too small; the other is there is no terminal state in this problem, except a few dangerous actions, the other actions share the similar rewards.
3. 3~5 iterations are required depends on the different initial policy.
4.

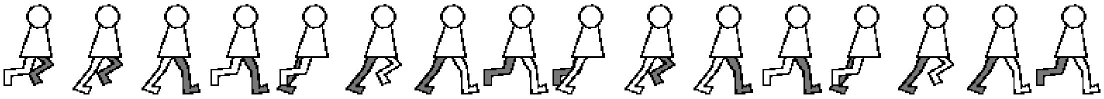


Figure 1: Policy iteration start from state 10

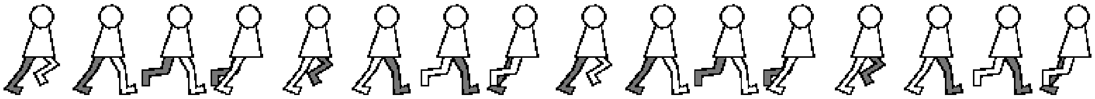


Figure 2: Policy iteration start from state 3

Q Learning

1. $\epsilon = 0.5$ and $\alpha = 0.8$ are adopted.
2. If a pure greedy policy is used, the optimal policy cannot be found. Because in this problem there is no terminal state, only a subset of states can be updated through a pure greedy policy.
3. Based on the parameters I chosen, it takes about 130 times for iteration.
4.

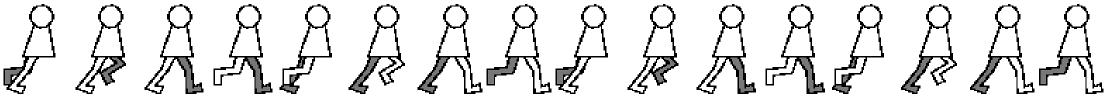


Figure 3: Q learning start from state 5

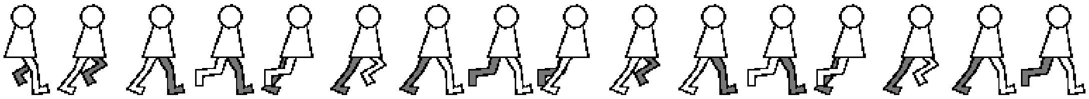


Figure 4: Q learning start from state 12