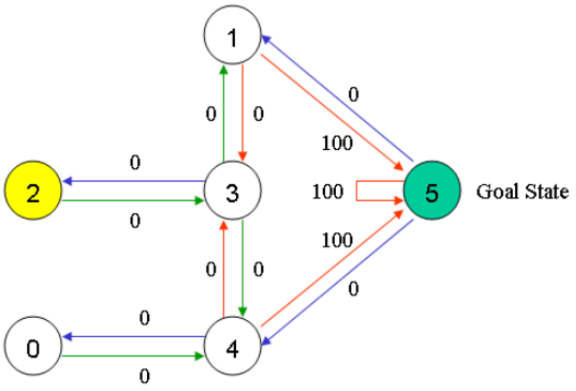
Homework 1

Reinforcement Learning : The Q-learning algorithm

Anna Vandi
Gabriele Cianni

Exercise 1



R-Matrix:

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	-1	-1	-1	-1	0	-1
1	-1	-1	-1	0	-1	100
2	-1	-1	-1	0	-1	-1
3	-1	0	-1	-1	0	-1
4	0	-1	-1	0	-1	100
5	-1	0	-1	-1	0	100

Q-Learning Algorithm

Question 1

Initialize the Q table (all 0)

State/Action	**Move to 0**	**Move to 1**	**Move to 2**	**Move to 3**	**Move to 4**	**Move to 5**
0	0	0	0	0	0	0
1	0	0	0	0	0	0

2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

Question 2

Episode 1: start state 1
Looking at R-Table I see that there are 2 possible actions if I am in state 1:

- Go to state 3
- Go to state 5

Random selection: I want to go to 5

a) Bellman equation: compute new value of Q(1,5)

Bellman equation should be:
delta_Q(state, action) = R(state, action) + Gamma * Max[Q(next state, all actions)]-Q(state,action)

newQ(state, action) = **Q(state, action)** + ALPHA * **delta_Q(state, action)**

but (ALPHA=1) I can summarize as follow:
Q(state, action) = R(state, action) + Gamma * Max[Q(next state, all actions)]

Q(1, 5) = R(1, 5) + 0.8 * Max[Q(5,1), Q(5,4), Q(5,5)]=100+0=**100**

Update Q-table

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	0	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

Question 3

Episode 2: start state 3
Initial state: State 3

action: go to 1 (random)
Q(3, 1) = R(3, 1) + 0.8 * Max[**Q(1,5)**, Q(1,3), Q(1,1)]= 0+0.8*(**100**)=80

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	0	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	**80**	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

State 1 is not the goal state, I continue in this episode.
Starting state: 1
action: go to 5

Q(1, 5) = R(1, 5) + 0.8 * Max[Q(5,1), Q(5,4), Q(5,5)]= 100+0=100
Goal state has been reached, episode ends.
Q-table at the end of this episode.

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	0	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	**80**	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

Question 4

Explore more episodes to find Q-table reaching convergence values.

Episode 3: start state 4
Initial state: State 4
action: go to 5
Q(4, 5) = R(4, 5) + 0.8 * Max[Q(5,1), Q(5,4), Q(5,5)]= 100+0,8*(**0**)=100
Goal state reached

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	0	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	**80**	0	0	0	0
4	**0**	0	0	0	0	**100**
5	0	0	0	0	0	0

Episode 4: start state 0
Initial state: State 0
action: go to 4
Q(0, 4) = R(0, 4) + 0.8 * Max[Q(4,5), Q(4,3), Q(4,0)]= 0+0,8*(**100**)=80

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	**80**	0	0	0	0
4	**0**	0	0	0	0	**100**
5	0	0	0	0	0	0

Current state: State 4
action: go to 3
Q(4, 3) = R(4, 3) + 0.8 * Max[**Q(3,1)** , Q(3,4), Q(3,2)]= 0+0.8*(**80**)=64

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	**80**	0	0	0	0
4	**0**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Current state: State 3
action: go to 4
Q(3,4) = R(3,4) + 0.8 * Max[Q(4,5), Q(4,3), Q(4,0)]= 0+0,8*(**64**)=51,2

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	0	0	**100**
2	0	0	0	0	0	0
3	0	**80**	0	0	**51,2**	0
4	**0**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Starting state: State 4
action: go to 5
Q(4, 5) = R(4, 5) + 0.8 * Max[Q(5,1), Q(5,4), Q(5,5)]= 100+0,8*(**0**)=100
Goal state reached

Episode 5: start state 2
Initial state: State 2
action: go to 3
Q(2, 3) = R(2, 3) + 0.8 * Max[**Q(3,1)** , Q(3,4), Q(3,2)]= 0+0,8*(**80**)=64

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	0	0	**100**
2	0	0	0	**64**	0	0
3	0	**80**	0	0	**51,2**	0
4	**0**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Current state: State 3
action: go to 2
Q(3,2) = R(3,2) + 0.8 * Max[Q(2,3)]= 0+0,8*(64)=**51,2**

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0

1	0	0	0	0	0	**100**
2	0	0	0	**64**	0	0
3	0	**80**	**51,2**	0	**51,2**	0
4	**0**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Current state: State 2
action: go to 3
Q(2, 3) = R(2, 3) + 0.8 * Max[**Q(3,1)** , Q(3,4), Q(3,2)]= 0+0,8*(**80**)=64

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	0	0	**100**
2	0	0	0	**64**	0	0
3	0	**80**	**51,2**	0	**51,2**	0
4	**0**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Current state: State 3
action: go to 1
Q(3,1) = R(3,1) + 0.8 * Max[**Q(3,1)** , Q(3,4), Q(3,2)]= 0+0,8*(80)=**64**

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	0	0	**100**
2	0	0	0	**64**	0	0
3	0	**64**	**51,2**	0	**51,2**	0
4	**0**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Current state: State 1
action: go to 3
Q(1,3) = R(1,3) + 0.8 * Max[Q(1,3), Q(1,5)]= 0+0,8*(100)=**80**

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	**80**	0	**100**
2	0	0	0	**64**	0	0
3	0	**80**	**51,2**	0	**51,2**	0
4	**0**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Current state: State 3
action: go to 4
Q(3,4) = R(3,4) + 0.8 * Max[Q(4,5), Q(4,3), Q(4,0)]= 0+0,8*(100)=**80**

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	**80**	0	**100**
2	0	0	0	**64**	0	0
3	0	**80**	**51,2**	0	**80**	0
4	**0**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Current state: State 4
action: go to 0
Q(4,0) = R(4,0) + 0.8 * Max[Q(0,4)]= 0+0,8*(80)=**64**

State/Action	Move to 0	Move to 1	Move to 2	Move to 3	Move to 4	Move to 5
0	0	0	0	0	**80**	0
1	0	0	0	**80**	0	**100**
2	0	0	0	**64**	0	0
3	0	**80**	**51,2**	0	**80**	0
4	**64**	0	0	**64**	0	**100**
5	0	0	0	0	0	0

Question 5

The best sequence is: 2 → 3 → 1 → 5.
This is because in the initial state the action with the highest Q-value from the initial state 2 is to move to 3 with the value of 64. Then, from state 3, the actions with the highest Q-value are “Move to 1” and “Move to 4”, because they have the same Q-value 80. In this case, it is indifferent which one to choose. Our agent choses to move to state 1 and in this state the action with the highest value is “move to 5”, with a Q-value of 100. Goal state is reached.