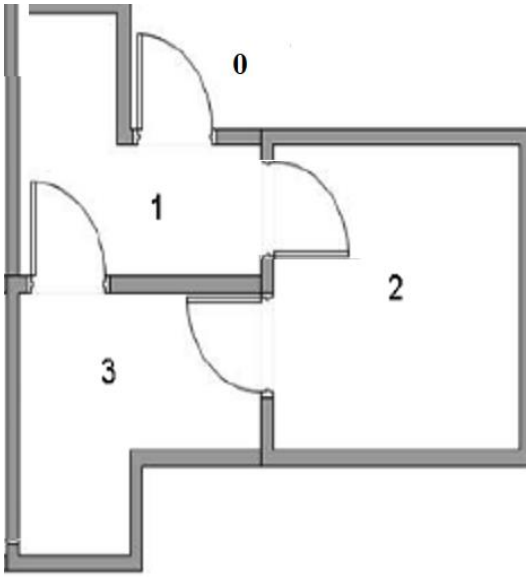
ECSE 4850/6850

Introduction to Deep Learning

Assignment #7

Due 11:59 pm, April 21th

1. Follow the room example in the lecture notes, perform Q learning using the deterministic model-free recursive equation to update the optimal Q function for the room configuration below. Use the same reward and Q function initializations. Note the goal is to find the optimal policy that produces the shortest path to go outside (room 0) from each room.



1) Identify the states, the actions, provide a generic reward function, and the initial Q function.
   4 states: 0, 1, 2, 3
   4 actions: 0-go outside, 1-go to room 1, 2-go to room 2, 3-go to room 3
   Reward function: 100 for outside move, -1 for illegal moves, 0 for other moves
   Reward table:

State\Action	0	1	2	3
0	100	0	-1	-1
1	100	-1	0	0
2	-1	0	-1	0
3	-1	0	0	-1

   Q function: all actions are 0
   Initial Q table:

State\Action	0	1	2	3
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0

2) Following the value-iteration pseudo code in the lecture notes to enumerate each state and action, show the Q function after each iteration, and produce the final learnt policy π with respect to each state, i.e., $a=\pi(s)$, in the form of the state-action table below.

State (s)	Action (a)
0	
1	
2	
3	

The value for gamma was not stated, so I will use .8 like in the lectures

Ignore illegal moves

Iteration 1:

With all Q-values = 0, the only values changing will be those that end with a reward

Q(0,0) = 100 + 0 = 100

Q(1,0) = 100 + 0 = 100

Q table

State\Action	0	1	2	3
0	100	0	0	0
1	100	0	0	0
2	0	0	0	0
3	0	0	0	0

Iteration 2:

Q(0,0) = 100 + .8*max{Q(0,1)} = 100 + .8*100 = 180

Q(0,1) = 0 + .8*100 = 80

Q(1,0) = 100 + .8*100 = 180

Q(2,1) = 0 + .8*100 = 80

Q(3,1) = 0 + .8*100 = 80

Q table

State\Action	0	1	2	3
0	180	80	0	0
1	180	0	0	0
2	0	80	0	0
3	0	80	0	0

Iteration 3:

Q(0,0) = 100 + .8*180 = 244

Q(0,1) = 0 + .8*180 = 144

Q(1,0) = 100 + .8*180 = 244

Q(1,2) = 0 + .8*80 = 64

Q(1,3) = 0 + .8*80 = 64

Q(2,1) = 0 + .8*180 = 144

Q(2,3) = 0 + .8*80 = 64

Q(3,1) = 0 + .8*180 = 144

Q(3,2) = 0 + .8*80 = 64

Q table

State\Action	0	1	2	3
0	244	144	0	0
1	244	0	64	64
2	0	144	0	64
3	0	144	64	0

Iteration 4:

Q(0,0) = 100 + .8*244 = 295.2

Q(0,1) = 0 + .8*244 = 195.2

Q(1,0) = 100 + .8*244 = 295.2

Q(1,2) = 0 + .8*144 = 115.2

Q(1,3) = 0 + .8*144 = 115.2

Q(2,1) = 0 + .8*244 = 195.2

Q(2,3) = 0 + .8*144 = 115.2

Q(3,1) = 0 + .8*244 = 195.2

Q(3,2) = 0 + .8*144 = 115.2

Q table

State\Action	0	1	2	3
0	295.2	195.2	0	0
1	295.2	0	115.2	115.2
2	0	195.2	0	115.2
3	0	195.2	115.2	0

Q values have converged