CS1571
Fall 2019
11/17 In-Class Worksheet

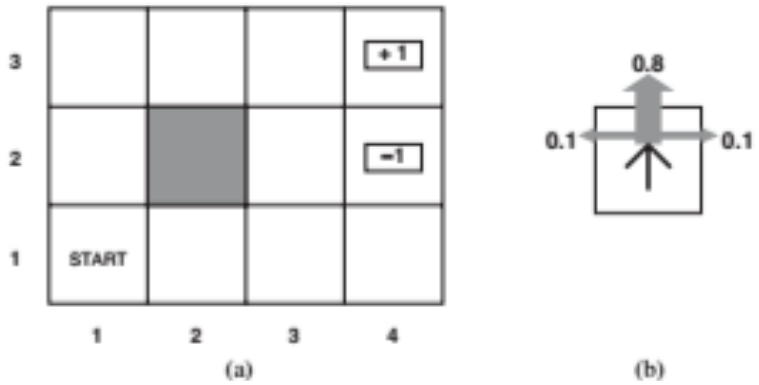
Name: Simran Gidwani
Where were you sitting in class today: Front right

Pre-Reflection

On a scale of 1-5, with 5 being most confident, how well do you think you could execute these learning objectives:

- 20.1 Define a Markov Decision Process. _____
- 20.2 Explain value iteration for MDPs. _____

A. Markov Decision Processes



1. Define this problem as a Markov Decision Process by states, actions, transition functions, and rewards. Note that +1 represents a desirable terminal state, and -1 represents a negative terminal state.

States: all the space coordinates on the grid
Actions: Move(x1, x2) where x2 is an adjacent square --- ex. Move((1, 1), (1, 2))

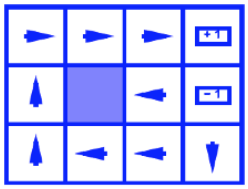
- Move forward, backward, left and right

Transition: either forward(.8) left(.10) or right square (.10)

- T((1, 1) move forward, (1, 2)) = .80
- P((1, 2) | P(1, 1), move forward) = .80

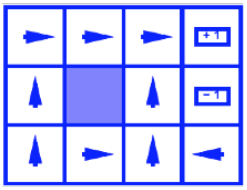
Rewards: R(4, 3) = +1
R(4, 2) = -1
R(x, y) = -.04

2. Notice that the optimal policy depends on how you conceptualize the rewards. Match the following rewards with the below policies.
    a. R(s) = -.01, R(s) = -0.03, R(s) = -0.4, R(s) = -2.0



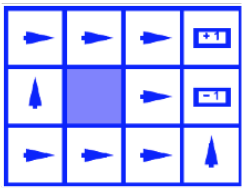
A

-.01



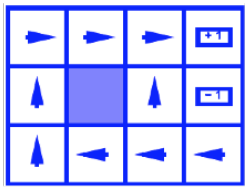
B

-.4



C

-2.0



D

-.03

Prioritizing finishing before getting to the best reward state

3. We control a robot who is traveling on a road that is five units wide; the robot is always moving forward; however, we must also always tell it to move either left or right. Whatever direction chosen, the robot has a 60% chance to move in that direction by 1 unit, 30% chance to move in that direction by 2 units, and 10% chance to not move in that direction at all. Dollar bills are placed along the far edges of the road. However, if the robot shoots off outside the road, the game immediately ends with a penalty of $100. We begin with the robot in the middle of the road. Although we could technically play this game forever, assume a discount constant of γ=0.8. Below is the problem set up with a MDP.

    States: offroad (either side), 1,2,3,4,5 (initial state is 3)
    Actions: left or right
    Transition T(s,a,s'):

	Next state					
State, action	off	1	2	3	4	5
1, left	0.9	0.1	0	0	0	0
1, right	0	0.1	0.6	0.3	0	0
2, left	0.3	0.6	0.1	0	0	0
2, right	0	0	0.1	0.6	0.3	0
3, left	0	0.3	0.6	0.1	0	0
3, right	0	0	0	0.1	0.6	0.3
4, left	0	0	0.3	0.6	0.1	0
4, right	0.3	0	0	0	0.1	0.6

5, left	0	0	0	0.3	0.6	0.1
5, right	0.9	0	0	0	0	0.1

Reward: associated with states:

state	off	1	2	3	4	5
Reward(state)	-100	1	0	0	0	1

Fill in the tables below to perform two iterations of value iteration:

$Q_1(s,a) = \sum_{s'} Pr(s'|s,a) [R(s')+\gamma V_0(s')]$

s	1		2		3		4		5	
a	left	right	left	right	left	right	left	right	left	right
Q(s,a)	-89.1	.1	-29.4	0	.3	.3	0	-29.4	.1	-89.1

Take the formula on the worksheet, for one left .9 * reward + gamma * vNot value

$V_1(s) = \max (Q_1(s, \text{left}), Q_1(s, \text{right}))$

s	off	1	2	3	4	5
$V_1(s)$	0*	.1	0	.3	0	.1

*since "offroad" is a terminal state, we can't choose any action for it. We'll just set V(offroad)=0
Max A—take whichever is larger Q(s, a)

$Q_2(s,a) = \sum_{s'} Pr(s'|s,a) [R(s')+\gamma V_1(s')]$

s	1		2		3		4		5	
a	left	right	left	right	left	right	left	right	left	right
$Q_2(s,a)$										

Value of .9 (ending up off the road) * reward of ending up off the road (-100) + gamma* prev utility + all other actions you can reach

$V_2(s) = \max (Q_2(s, \text{left}), Q_2(s, \text{right}))$

s	off	1	2	3	4	5
$V_2(s)$	0					

What is the policy after two iterations (what actions should be performed in each state)?

s	off	1	2	3	4	5

π(s)	-					

B. Post-Reflection

On a scale of 1-5, with 5 being most confident, how well do you think you could execute these learning objectives:

- 20.1 Define a Markov Decision Process. _____
- 20.2 Explain value iteration for MDPs. _____