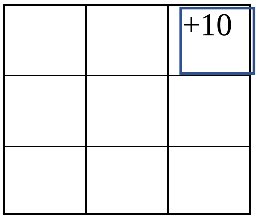
C S 4013/5013: Artificial Intelligence
Spring 2022
University of Oklahoma
Homework Assignment 7, Due: 4/29/22, 11:59 PM

Question 1 (MDP): [50 points]
Consider the 3 × 3 world shown in the figure below. The state indicated by square is the terminal state. The transition model is as follows: 80% of the time the agent goes in the direction it selects; the rest of the time it moves at right angles to the intended direction (either move to the left or right). Note that if it hits the borders, it will stay in the current state.

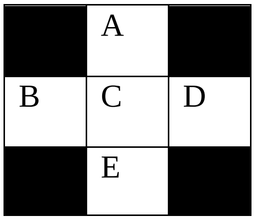
(a) Implement value iteration for this world for each value of living (I.e., transition or temporary) reward r below for k=5 iteration. Use discounted rewards with a discount factor of 0.99. Show the policy obtained in each case. Explain intuitively why the value of reward r leads to each policy.

r = 10
r = 0
r = -100



(b) Implement policy iteration. Assume that the initial policy π_i is moving to the east, except the terminal state that the policy is to exit. Complete the following values after one iteration of policy iteration. Run value iteration for k=2 (you may write a simple program to compute it)
$V_k^{\pi_i}$ =?
π_{i+1} =? (updated policy after one iteration)

Question 2 (RL) [50 points]:
Consider the following grid world with five different states. The actions are move east, west, south, north, and exit if it is in a terminal state.



(a) We would like to use Model-based learning using the following four observations. What is the estimated Transition and reward based on these observations?

Episode 1

B, east, C, -1
C, east, A, -1
A, exit, x, -5

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, x, +5

Episode 3

E, north, C, -1
C, east, D, -1
D, exit, x, +5

Episode 4

E, north, C, -1
C, east, A, -1
A, exit, x, -5

(b) Implement direct evaluation as a model-free based learning based on those four observations and calculate the value states for each state. Assume $\gamma = 1$.

(c) We would like to use TD learning and Q-learning to find the values of these states. Suppose that we have the following observed transitions:

(B, East, C,3), (C, South, E, 2), (C, East, E,3) , (D, West, C,2), (A,South,C,3)

The initial value of each state is 0. Assume that $\gamma = 1$ and $\alpha = 0.5$.

What are the learned values from TD learning after all five observations? Show the process of computing these values.

(d) What are the learned Q-values from Q-learning after all five observations? Show the process of computing these values.