CSE 574

HW 4

Due Wednesday Nov 7 at 11:59pm

*Please submit on Blackboard. Show work for full credit. See the class policies for late HW assignments as we will strictly enforce them.*

Name:

ASU ID:

1. **Passive Reinforcement Learning**

Given the 4x3 grid world below, initialize all the utility values to 0. Set the utility of (4,3) to 100 and (4,2) to -100,

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | 100 |
|  |  |  | -100 |
|  |  |  |  |

1 2 3 4

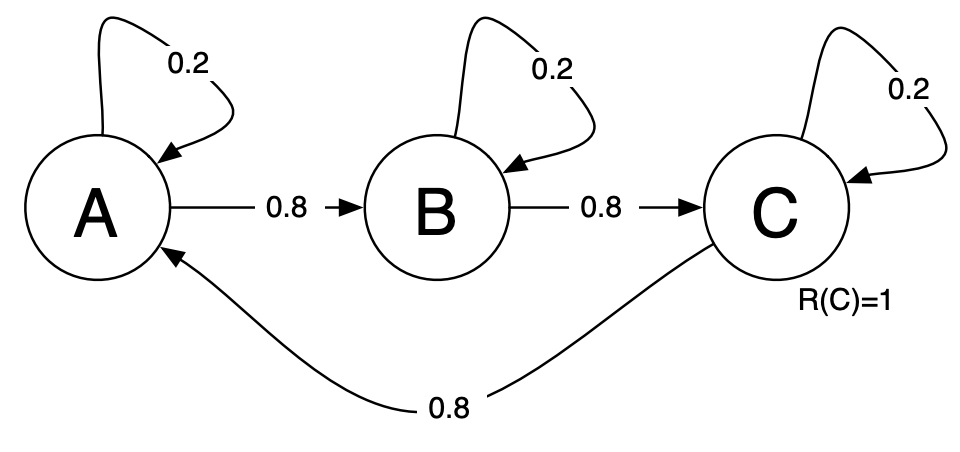
1. Instead of setting reward function for each block with fixed value in lecture, let’s define the R(n) = -n2. Let n denotes the number of steps from beginning, so the reward value for (1,1) is 1. Given two trials starts from (1,1):
2. (1,1)->(1,2)->(1,3)->(2,3)->(1,3)->(2,3)->(3,3)->(4,3)
3. (1,1)->(2,1)->(3,1)->(3,2)->(4,2)

Apply direct utility and adaptive dynamic programming separately, update the value for each visited grid, show the work.

1. Let R(n) to be (n-)2, apply the direct utility method on both trials, compute the which minimize the utility of (1,1).

2. **Adaptive Dynamic Programming**

Given an MDP with states transition graph below:



For each state, the agent could have two operations:

Move: move to next state with probability 0.8, and stay with probability 0.2.

Stay: 100% stays in current state.

Initially, the reward for C is 1 otherwise 0. Let the discount factor to be 0.9 and learning rate to be 1.

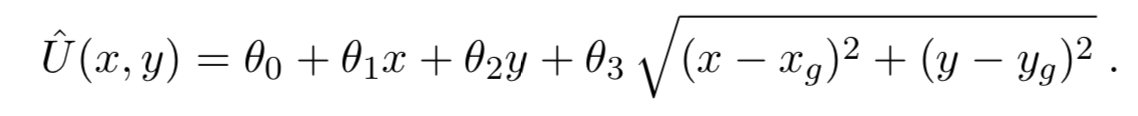
1. Suppose we have a trial:

(Move ,A)->(Move, B)->(Move, C), please show the Q value for each state-action pair.

1. Think about apply adaptive dynamic programming, which method would be better if we let the chain to be very long?
2. Based on last question, what would be disadvantages of ADP compared to Q learning?

3. **Active Reinforcement Learning**

Write out the parameter update equations for TD learning with

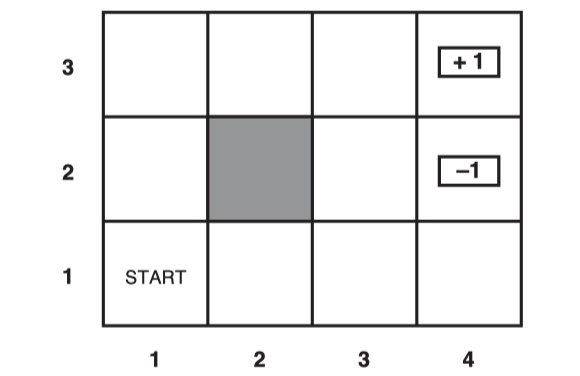


4. **Partially Observable Markov Decision Process**

POMDP

Ref: Norvigs’s textbook, page 658 and page 691 Problem 17.13

1. Let’s convert the 4 × 3 world of Figure below into a POMDP by adding a noisy or partial sensor instead of assuming that the agent knows its location exactly. Such a sensor might measure the number of adjacent walls, which happens to be 2 in all the nonterminal squares except for those in the third column, where the value is 1; a noisy version might give the wrong value with probability 0.1. Let the initial belief state b 0 for the 4 × 3 POMDP on page 658 be the uniform distribution over the nonterminal states, i.e., Calculate the exact belief state b 1 after the agent moves Left and its sensor reports 1 adjacent wall.

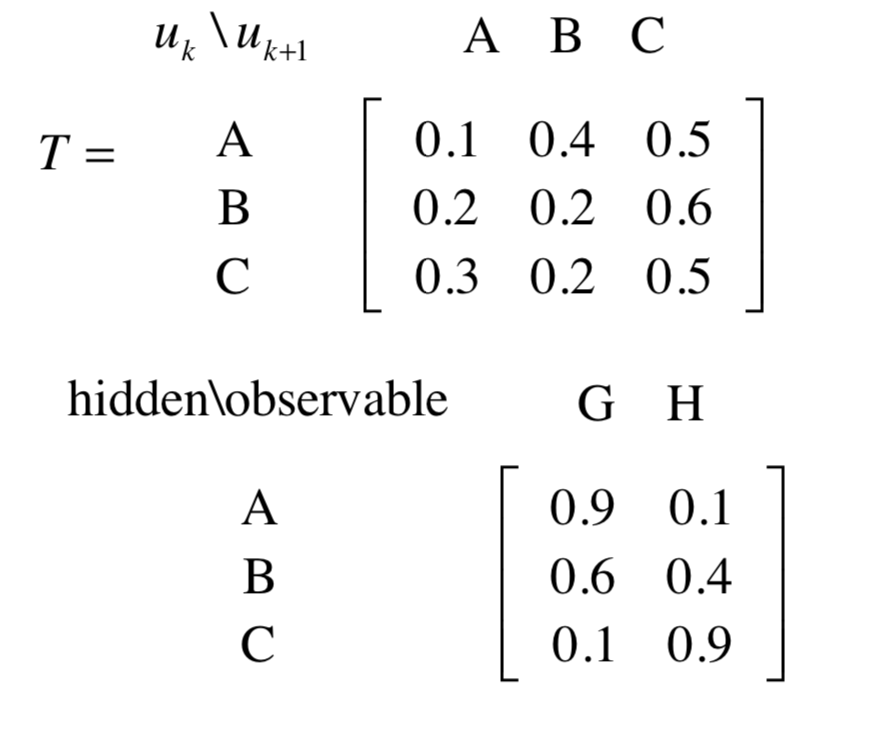


1. What is the time complexity of d steps of POMDP value iteration for a sensorless environment?

5. **Hidden Markov Models**

Given below the state transition probabilities and observation probability from each state

States S = {A, B, C} and observations O = {G, H}, Initial probabilities π(A) = 0.2, π(B) = 0.1, π(C) = 0.7



a) Draw the hidden Markov representation

b) Find the maximum joint probability P(u, HHGH) for hidden state sequence u.

c) Calculate the maximum conditional probability P(u | HHGH) for hidden state sequence u.

d) Give the hidden state sequence u that achieves these maximum.