Assignment 12

Reinforcement Learning

Prof. B. Ravindran

Instructions: In the following questions, one or more choices may be correct. Select all that apply.

1. Consider an environment in which an agent is randomly dropped into either state s_1 or s_2 with equal probability. The agent can only view obstacles present immediately to the North, South, East or West. However the observation made in each direction by the agent may be wrong with a probability of 0.1. If in state s_1 obstacles are present to the North and South, and in s_2 obstacles are present to the East and West, what is the probability of the agent being in state s_1 if the observation made is that there are obstacles present to the North and West.

   (a) 81/82

   (b) 41/82

   (c) 73/82

   (d) None of the above.

   Sol. (b)
   Application of Bayes Rule.$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\neg A)P(\neg A)}$ with A being the "Present in state s_1" and B being "Observed obstacles to the North and West. Ans = 0.5"

2. In the same environment as Question 1, suppose state s_1 has obstacles present only to the North and South, and s_2 has obstacles present only to the East and West. What is the probability of the agent being in state s_1 if the observation made is that there are obstacles present only to the North, East and West.

   (a) 81/82

   (b) 41/82

   (c) 73/82

   (d) None of the above.

   Sol. (d)
   Application of Bayes Rule.$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\neg A)P(\neg A)}$ with A being the "Present in state s_1" and B being "Observed obstacles to the North, East. Ans: 1/82"

3. **Assertion:** One of the reasons that history-based methods are not feasible is because of the significant increase in the state space size when the trajectory is long.
   Reason: The number of states increases polynomially w.r.t. trajectory length.

   (a) Both Assertion and Reason are true, and Reason is correct explanation for Assertion.

   (b) Both Assertion and Reason are true, but Reason is not correct explanation for assertion.

   (c) Assertion is true, Reason is false

   (d) Both Assertion and Reason are false

Sol. (c)
The assertion is true, the number of states increases significantly with trajectory length, however it is due to the exponential increase in the number of states that history-based methods are not feasible.

4. In the case of POMDPs, which of the following is a good estimate of the return of a trajectory of a policy π, given the current belief state and the solution to the underlying MDP (value function for all states)?

   (a) Average of all $V^{\pi}(s)$ where $b(s) > 0$.

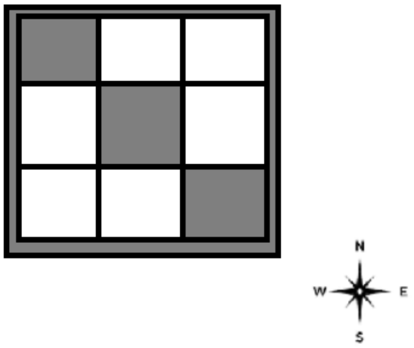
   (b) Weighted average of all $V^{\pi}(s)$ where $b(s)$ are the weights.

   (c) Average of all $V^{\pi}(s)$ where $b(s) \geq \alpha$, where α is a small positive value

   (d) None of the above

   Sol. (b)
   It is equivalent to the expected value of the return, making it a good estimate of the return of the current trajectory.

5. Consider the grid-world shown below:



   Walls and obstacles are colored gray. The agent is equipped with a sensor that can detect the presence of walls or obstacles immediately to its North, South, East or West.

   Which of the following are **true** if we represent states by their sensor observations?

   (a) The grid-world is a 1st-order Markov system.

   (b) The grid-world is a 2nd-order Markov system.

   (c) The grid-world is a 3rd-order Markov system.

   (d) The grid-world is a 4th-order Markov system.

   Sol. (a),(b),(c),(d)
   Each state is uniquely identifiable from the sensor reading. A system that is 1st order Markov is necessarily 2nd, 3rd and 4th order Markov.