National University of Singapore
School of Computing

Semester 1, AY2023-24

CS4246/CS5446 AI Planning and Decision Making

Tutorial Week 10: POMDP

Guidelines

You may discuss the content of the questions with your classmates. But everyone should work on and be ready to present ALL the solutions.

Problem 1: POMDP Solving

[Modified from RN 17.13] We can convert the 4×3 world we have seen in the lectures into a POMDP by adding a noisy 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 be b_0 for the 4×3 POMDP on page 658 be the uniform distribution over the non-terminal states, i.e.,

$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	0
$\frac{1}{9}$	$\times$	$\frac{1}{9}$	0
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

Calculate the exact belief state b_1 (rounded off to 5 decimal places) after the agent moves *Left* and its sensor reports 1 adjacent wall.

Solution:

We want to calculate the new belief state b_1, given the *Left* action and the sensor seeing 1 wall. We will calculate this in parts, and then normalize the final results. First, we want to calculate the probability $P(x'|Left, b_0) = \sum_x P(x'|Left, x)b_0(x)$ of reaching each state x' after the *Left* action:

0.2	$\frac{1}{9}$	$\frac{0.2}{9}$	0
$\frac{1}{9}$	$\times$	$\frac{1}{9}$	$\frac{0.1}{9}$
0.2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{0.1}{9}$

Now, we update these estimates with the sensor data, which says there is one adjacent wall (i.e.,

multiply by $P(z = \text{`1 adjacent wall'}|x')$):

0.1×0.2	$0.1 \times \frac{1}{9}$	$0.9 \times \frac{0.2}{9}$	0
$0.1 \times \frac{1}{9}$	$\times$	$0.9 \times \frac{1}{9}$	$0.9 \times \frac{0.1}{9}$
0.1×0.2	$0.1 \times \frac{1}{9}$	$0.9 \times \frac{1}{9}$	$0.1 \times \frac{0.1}{9}$

and renormalize to get b_1:

0.06569	0.03650	0.06569	0
0.03650	$\times$	0.32847	0.03285
0.06569	0.03650	0.32847	0.00365

Problem 2: Modeling with POMDPs

1. Dialog system. A common form of a dialogue system is a slot-filling dialog system, where there the user's goals is defined by a set of variables and the aim of the dialogue is to elicit the values of the variables from the user. For example, in an automated taxi booking system, the system may try to find out the destination location and the pick-up location from the user.

   Solution:
   Each assignment of values to the variables $(v_1, \ldots, v_n)$ is a possible state. An action may be to ask for the value of a particular variable. Another possible action is to ask for confirmation (yes or no). A reasonable assumption is that the user does not change his or her mind, so the true values of the state does not change, i.e. a state transitions back to itself. The observation may depend on the question asked, if asked for possible value of the variable, the observation are the possible values. If asked for confirmation, the observation is "yes" or "no". The observation function is the probability of the value being received, given the true state of the variable being asked for questions asking about a variable. For confirmation questions, the observation function is on receiving "yes" or "no" given the question and true state of the variable being asked. One additional action would be to terminate the conversation and execute the outcome of the dialog, e.g. dispatch a taxi. A reasonable reward to be to give a good reward for the correct action and a penalty for an incorrect action. In addition, it would likely be useful to penalize each question asked (possibly differently for different types of questions) in order to encourage the agent to minimize the number of questions asked.

2. Kermit the Frog is stuck at a corner of a lotus pond. There are only two lotus leaves at the corner, one Red and one Green, and staying on Green long enough (to bend the stem to open a path) will lead him out of the corner. But poor Kermit is color-blind and cannot tell exactly which leaf he is on; he has also hurt his left leg. When Kermit jumps, there is 80% chance that he will actually jump to another leaf. Also, Kermit can estimate his position (on Red or Green) from the texture of the leaf with 70% accuracy.

   Stating any additional assumptions you may need, formulate Kermit's problem as a sequential decision problem. Carefully list out all the components of the model.

Solution:

– Two states: A and B (Red and Green) (Belief space is 1-dimensional)

– Rewards: $R(A) = 0$ and $R(B) = 1$

– Two actions with transition model:

1- Stay – Stays put with probability = 1

2- Jump – Jumps to the other state with probability = 0.8

– Transition Function:

$T(A, 1, A) = 1$
$T(A, 1, B) = 0$
$T(A, 2, A) = 0.2$
$T(A, 2, B) = 0.8$
$T(B, 1, A) = 0$
$T(B, 1, B) = 1$
$T(B, 2, A) = 0.8$
$T(B, 2, B) = 0.2$

– Evidence Space: $\{C, \neg C\}$ – Correct or incorrect Observation

– Sensor with observation model: Reports correct state with probability = 0.7

– Observation function:

$O(A, C) = 0.7$
$O(A, \neg C) = 0.3$
$O(B, C) = 0.7$
$O(B, \neg C) = 0.3$

Problem 3: Captain Jack's Adventure

Captain Jack would like to go to Treasure Island (Island) but does not know the way. He knows that the Island is on his left (state s_1) with probability p and the Island is on his right (state s_2) with probability $1 - p$. If he goes in the wrong direction, he would end up in Pirates Den (Den), a place that he wants to avoid badly. Captain Jack has three possible actions. He can go left (action a_1), go right (action a_2), or ask the Lighthouse Keeper (Keeper) at his current docking harbor (action a_3) whether to go left or right. If he goes in the correct direction, he gets a reward of 100 (e.g. $R(s_1, a_1) = 100$) but if he goes in the wrong direction he gets a penalty of -100 (e.g. $R(s_1, a_2) = -100$). The Keeper never lies, providing the observations left for Island on the left, and right for Island on the right. But asking the Keeper will cost -10 (i.e. $R(s_1, a_3) = R(s_2, a_3) = -10$).

(a) The value of a one-step plan taken in state s is simply the reward of taking the action a in state s: $R(s, a)$. Going left or right are terminal actions while asking the Keeper is non-terminal. Hence, two-step conditional plans can only start with the non-terminal action of asking the Keeper (a_3) followed by an observation and ends with taking another action.

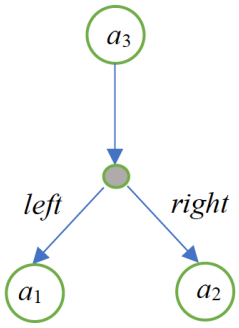
(i) How many two-step conditional plans that starts with action a_3 are there?

(ii) There is only one non-dominated two-step conditional plan: draw (or clearly describe) the non-dominated two step conditional plan.

Solution:

(i) In the policy tree, there are three possible action choices (go left, go right, ask Keeper) under the observation left and three possible action choices under observation right giving 9 possible two-step conditional plans. One such policy tree is shown in (ii) – there are 9 such trees.

(ii) The non-dominated two-step conditional plan is shown below.



(b) The one-step plan consisting of asking the Keeper cannot be optimal. Hence there can be at most two non-dominated one-step plans. From part (a) of this question, we know that there is only one non-dominated two-step conditional plan, giving a total of 3 non-dominated one and two step plans.

(i) Give the three α-vectors corresponding to the three non-dominated plans. Assume that the discount factor is $\gamma = 1$ (not discounted).

Solution:
Action left: $\alpha_l(s_1) = R(s_1, a_1) = 100$, $\alpha_l(s_2) = R(s_2, a_1) = -100$
Action right: $\alpha_r(s_1) = R(s_1, a_2) = -100$, $\alpha_l(s_2) = R(s_2, a_2) = 100$
Two-step plan: $\alpha_p(s_1) = \alpha_p(s_2) = -10 + 100 = 90$

(ii) Partition the beliefs into regions where each plan is optimal. Describe the regions.

Solution:
The region where action left is optimal is $100p - 100(1 - p) \geq 90$ giving the region

$p \geq 19/20$. The regions where action right is optimal is $-100p + 100(1-p) \geq 90$ giving the region $p \leq 1/20$. In the remaining region $1/20 \leq p \leq 19/20$, the two-step plan is optimal.
