Assignment 12 (Sol.)

Reinforcement Learning

Prof. B. Ravindran

1. In the partial observability example that we saw, there were some observations which uniquely identify the state of the environment. Suppose in a problem, no such observations exist, i.e., there are no observations which allow us to exactly determine the state of the environment. In such problems, is it ever possible to be sure which state you are in given only the sequence of observations?

   (a) no

   (b) yes

   Sol. (b)
   It is not necessary that there exist observations that uniquely identify the state of the system. In many problems, the sequence of observations suffice to determine with certainty, the state of the system. For example, in a problem, a particular observation may correspond to multiple states. However, it is possible that the sequence of observations that precede such an observation allows us to distinguish between the multiple states which that observation corresponds to. In addition, the agent has knowledge of its actions which can further help make such distinctions.

2. Suppose that some of the actions that are available to an agent in a partially observable environment have non-deterministic outcomes. How would this impact the agent's ability to determine the state it is in?

   (a) it would make determining the state easier

   (b) it would make determining the state harder

   (c) it would have no effect in determining the state

   Sol. (b)
   Non-deterministic action outcomes introduce another layer of uncertainty, since the agent can now not be sure what the effect of taking an action in a state is (as the observations do not allow the agent to deterministically evaluate the state to which the agent transitions to after taking an action). This would generally make the problem of state determination harder.

3. Referring to the partial observability example considered in the lectures and assuming that there is no noise in the sensor outputs, is it true that all observation sequences of length 3 or more result in the elimination of uncertainty regarding the position of the agent in the environment?

(a) no

(b) yes

Sol. (a)
Consider the plausible sequence of observations 1010, 1010, 1010, 1010.

4. To solve a POMDP problem, suppose you decide to maintain belief states. Would the estimation of the belief states benefit from access to the history of observations and/or actions?

   (a) access to past observations, but not actions, would improve belief state estimation

   (b) access to past actions, but not observations, would improve belief state estimation

   (c) access to past observations and actions would improve belief state estimation

   (d) access to past observations or actions will not improve belief state estimation

   Sol. (d)
   An agent needs to update its belief upon taking an action and observing the corresponding observation. Since the state is Markovian, maintaining a belief over the states solely requires knowledge of the previous belief state, the action taken, and the current observation. The information from past actions and observations is already encapsulated in the belief state.

5. Suppose that you are given a problem in which the agent is able to determine the state it transitions to after each action taken by the agent. However, the state space is described in a manner which requires the agent to consider information in the current, as well as the latest past state in the sequence, to make action decisions. Is the problem so defined an MDP, a POMDP, or neither? Can this problem be solved using RL techniques that we have studied (perhaps with possible modifications to the problem definition)?

   (a) MDP, yes

   (b) POMDP, yes

   (c) neither, no

   (d) neither, yes

   Sol. (d)
   The problem so defined is neither an MDP (the Markov property does not hold) nor a POMDP (partial observability is not an issue here). We can reformulate the problem by considering a state in the new formulation to be a subset of the set of all pairs of states in the original formulation. With appropriate modifications to the transition probability and reward function, we can obtain a MDP which can be solved using the various RL techniques that we have seen. Another approach would be to consider this problem as a POMDP where each 'observation' gives only partial information regarding the 'state' of the system and use history based methods.

6. Recall the example partially observable environment encountered in the lectures. Suppose the goal in this problem is to reach the bottom-right state, with the usual -1 rewards for each transition. Assume that the agent starts two cells below the top left state and the first action selected by the QMDP procedure is to move down. Assuming deterministic action outcomes with no noise in the observations, as well as the use of the QMDP procedure where the underlying MDP has been correctly solved, what can you say about the optimality of the agent's policy in this scenario?

(a) the agent will follow an optimal policy to reach the goal

(b) the agent's policy to reach the goal will not be optimal

(c) the optimality of the agent's policy towards reaching the goal in this scenario cannot be determined with certainty

Sol. (a)
Note that on taking its first action, the agent reaches a state which is uniquely identifiable by the the observation corresponding to that state. Thus, the first action of the agent dispels all state uncertainty and this allows the QMDP procedure to follow the optimal policy to reach the goal.