CSE 571: Artificial Intelligence

[Fall 2009] 11/9/2009

Mid-semester Sanity Test

Your Name:_____

Instructions: This test is closed book and closed notes. It has to be completed in class. **Please limit your answers the space provided.**

Make sure you take a quick look at the entire test before answering any part. Don't get stuck in any single part; there may be other parts that you know the answers for.

List of Questions:

I. Short answer questions

II. Numbers/Numbers/Numbers

III. True/False Questions

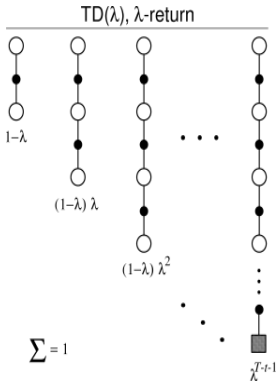
IV. Think Belief Search

V. Think Q-learning

VI. The "*Aw C'mon you barely even touched on any of the things I did learn*" question

I . **Short Answer Questions:**

1. Why is the weight for the last backup of TD(λ) not in the same form as the remaining ones? How does this show that it is a middle ground between TD learning and Monte Carlo learning?



2. Given an MDP, with transition function T(S,a,S'); and reward function R(S), write down the Bellman equation for the optimal value of state S (assume infinite horizon; either discounted or undiscounted):

V*(S) =

Given the V*(.) function for the MDP, explain how you can compute π*(S) (where π* is the optimal policy)

π*(S) =

3. Why is the straightforward application of *likelihood weighting* approach (for approximate inference) problematic for dynamic bayes nets? What is the alternative idea and how does it avoid the problem?

4. What is *ergodicity* and why is it important for online search?

5. In a belief space search, when a non-deterministic action **A** is applied to a belief state **B** giving rise to **B'**. Explain when (i.e., for what kind of actions) the cardinality of **B'** be (a) equal to (b) less than and (c) greater than the cardinality of **B.**

6. Explain why the state estimation problem is more interesting in stochastic worlds rather than non-deterministic worlds.

7. In discussing policy gradient search, we saw that it uses a mixed policy rather than a deterministic policy. What is the motivation for this?

8. Give an example of conditional preference independence relation (use any common sense domain).

9. How does the GAI (Generalized Additive Independence) model cover both the case of (a) full independence and (b) full dependence among the utilities of attributes? How many numbers ("parameters") are needed to specify it in each case?

10. For stochastic planning, heuristics can be devised by either relaxing uncertainty, or relaxing subgoal interactions or both. In each of the following cases, explain what relaxations are used: (a) FF-HOP (b) POND/McLUG.

II[Numbers, Numbers]:

Consider an agent that inhabits a world where states are described in terms of *n* Boolean state variables. At each state, the agent can do any of *A* actions which may be *deterministic, non-deterministic or stochastic*.

1. How many states are there in the agent's world?

2. How many belief states are there in the agent's world if none of its actions are stochastic? What if the actions are stochastic?

3. How many distinct deterministic policies are there in the agent's world? Does this depend on whether the actions are non-deterministic or stochastic?

4. How many *mixed policies* are there in the agent's world?

Assume now that the actions are all stochastic (with deterministic being a special case).

5. What is the size of the agent's value function in the extensional (tabular) representation?

6. What is the size of the agent's Q-value function?

Now let's cracking with some utilities.

7. What is the maximum number of independent utility values that need to be specified for this world? What is the minimum number (assuming the states don't all have the same utility).

III . True/False Questions

For the following, decide whether the statement is true or false, and **justify** your answer briefly.

1. Because it is model-free, a Q-learning agent is like a "hill-climbing" agent in that it requires no memory.

2. If the FF planner was optimal, then FF-replan planer using the *most likely outcome* determinization will be admissible.

3. Online search is needed only when the agent doesn't have a good model of its environment.

4. Factored approaches for reinforcement learning are motivated solely for reducing the memory requirements.

5. The cardinality heuristic used by conformant planners can also be gainfully adapted to classical planning.

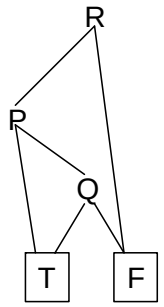
6. The fact that the utility of the expected monetary value of a lottery is higher than the lottery itself shows that in uncertain environments, the agent not only needs to provide utilities for each state but also for each lottery.

7. The effectiveness of the label propagation procedure of the LRTDP algorithm reduces as the world contains more and more sink states.

8. Since MDPs only consider causative actions, and do not have observational actions, they are in essence aimed at non-observable environments.

9. Preference compilation process may introduce preferences between outcomes for which the agent originally didn't express any preferences.

IV . Think Belief States



Consider a belief state B represented by the BDD shown on the right in a world where states are described in terms of 3 boolean state variables P, Q and R.

4.1 List the set of complete states represented by the BDD.

We have an action **A** which has two effects:

(i) If P is true, make P false.

(ii) If Q is true, make Q false.

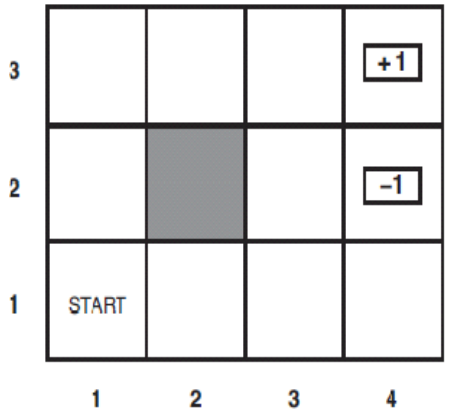
4.2. What is the result when **A** is applied to **B**?[It is fine to just list the states in the resulting belief state. Bonus if you can write the BDD].

4.3 We have an observational action **O** which tells us whether P is true or false.

What is the result when **O** is applied to **B**?

4.4. The agent's goal is the belief state {(~P, Q, R)}. Is the goal satisfied in the belief state resulting in part 4.2? How about after part 4.3?

V. Think Q-Learning



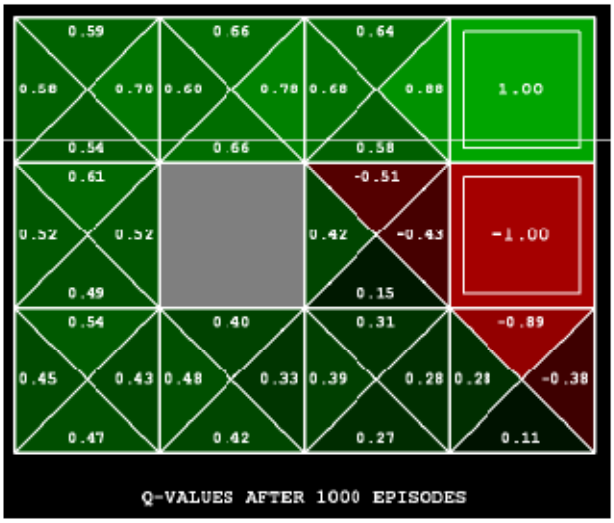
Consider a grid world shown on the right, where the cell (1,1) is the start state, while the cell (4,3) is a terminal state with reward +1 and (4,2) is a terminal state with reward -1. Every other state has a small negative reward of -0.04. The agent has four actions, UP, DOWN, LEFT, and RIGHT, and initially it doesn't know their model.

Assume that initially, all the values and q-values are set to 0 (except for the values of terminal states, which are set to their immediate reward values).

The agent starts from the start state (1,1) and does the following actions giving rise to the following states, and ending in a terminal state.

(1,1)→UP→(1,2)→LEFT→(1,3)→RIGHT→(2,3)→RIGHT→(3,3)→DOWN→(4,2)

5.1 How does Q-learning update **Q((3,3), DOWN)** in this trace? (Assume the learning rate, alpha, is 0.1; and gamma the discount rate is 0.9).



The figure on the right shows the Q-values after some 1000 iterations of Q-learning.

5.2 What is the value of the state (2,1)? What is the recommended action? Can you recover the "model" of the action from the Q-values?(*if the print out is illegible, the numbers in the cell (2,1), in clockwise direction starting from bottom, are 0.42, 0.48, 0.40 and 0.33).*

VI. The "**Aw, C'mon. You barely even touched on any of the things I did learn"** question

List five (5) non-trivial technical ideas you gleaned from the class up to this point:

1

2

3

4

5