

CPSC 422

Practice Midterm Exam Questions

March 2007

This is meant to give some example questions. This is much longer than the exam. The exam may be nothing like this. But if you can do this, you should have no problems with the exam.

You may bring in one letter sized piece of paper with anything written on it. You may not use calculators, PDAs, robotic assistants or other electronic aids.

Some important points (that students often forget):

- Read and answer the question. You will not get marks for writing things (whether they are true or not) that are not relevant to the question.
- Use proper English in full sentences. You will not get marks if we cannot work out what you are saying.
- If a question asks about a particular instance of a problem, make sure your answer refers to that instance. Writing a general formula that you may have copied from the sheet you can bring in, is not worth any marks. (The questions are usually asking to apply that formula to a particular case, to make sure you understand it).

1. Answer the following questions. Use proper English. Be concise in your answers. (You will lose marks for stating irrelevant facts). You must use your own words (text from the textbook or another source copied onto your crib sheet will not get any marks).
 - (a) Give the intuition behind the notion of a transduction? Why do we define causal transductions?
 - (b) Why does an agent have a belief state?
 - (c) Explain why we need a command function and a state transition function in a robot controller but not other functions.

2. Suppose our Q-learning agent, with fixed α , and discount gamma, was in state 34 did action 7, received reward 3 and ended up in state 65. What value(s) get updated? Give an expression for the new value. (You need to be as specific as possible)

Solution $q[34, 7] = q[34, 7] + \alpha * (3 + \text{gamma} * \max_a q[65, a] - q[34, 7])$

3. In temporal difference learning (e.g. Q-learning), to get the average of a sequence of k values, we let $\alpha_k = 1/k$. Explain why it may be advantageous to keep α_k fixed in the context of reinforcement learning.

Solution The initial values are not as good estimates as newer values, and so you may not want to weight them as much. It is simpler to ignore the counts (and so keep α fixed). With a fixed α it is able to adjust when the environment changes.

4. Explain what happens in reinforcement learning if the agent always chooses the action that maximizes the Q-value. Suggest two ways that can force the agent to explore.

Solution It gets stuck in non-optimal policies because it does not explore enough to find the best action from each state. To explore, it can pick random actions occasionally. You could also set the initial values high, so that unexplored regions look good.

5. In MDPs and reinforcement learning explain why we often use discounting of future rewards.

Solution With no discounting the sum of the rewards is often infinite. Discounting means that more recent rewards are more valuable than rewards far in the future.

6. What is the main difference between asynchronous value iteration and standard value iteration? Why does asynchronous value iteration often work better than standard value iteration?

Solution In standard value iteration all of the values are updated from the previous values in a sweep through the values. In asynchronous value iteration, the values are updated from the current value and can be done in any order (you don't need to sweep through all of the values). It often works better because the latest values are always used and it can concentrate on updating values where they make the most difference (as it doesn't need to sweep through all of the values each time).

7. What is the relationship between asynchronous value iteration and Q-learning.

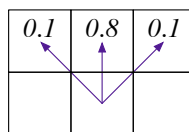
Solution Q-learning is like asynchronous value iteration in that it updates the Q values, but it uses experience rather than using a model. Thus the average is the average over its experience rather than computing the expected value using a model.

8. In feature-based reinforcement learning, what new data point does the experience $\langle s, a, r, s', a' \rangle$ give for the linear regression? (I.e., what is the new data point for what value?)

Solution It gives a new estimate of $r + \gamma Q(s', a')$ for the value of $Q(s, a)$.

9. Consider a 5×5 grid game similar to the simple game used in assignments. The agent can be at one of the 25 locations, and there can be a treasure at one of the corners or no treasure.

In this game the “up” action has the dynamics given by:

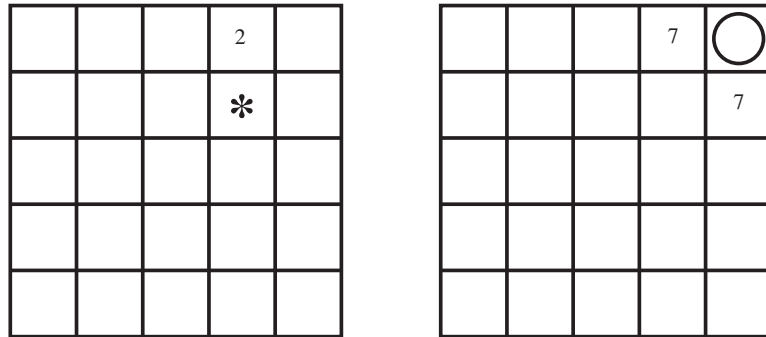


That is the agent goes up with probability 0.8 and goes up-left with probability 0.1 and up-right with probability 0.1.

If there is no treasure, a treasure can appear with probability 0.2. When it appears, it appears randomly at one of the corners, and each corner has an equal probability of treasure appearing. The treasure stays where it is until the agent lands on the square where the treasure is, and the agent gets an immediate reward of +10, and the treasure disappears in the next state transition. The agent and the treasure move simultaneously so that if the

agent arrives at a square at the same time the treasure appears at the same time, it gets the reward.

Suppose we are doing asynchronous value iteration and have the following value for each state:



where the left grid shows the values for the states where there is no treasure and the right grid shows the values of the states when there is a treasure at the top-right.

Consider the next step of asynchronous value iteration. For state s_{13} , which is marked by $*$ in the above figure, and the action a_2 which is “up”, what value is assigned to $Q[s_{13}, a_2]$ on the next iteration of value iteration? You need to show all working, but don’t need to do any arithmetic (i.e., leave it as an expression). Explain each terms in your expression.

Solution There are 15 possible states that could be entered, depending on which direction the robot actually went (up, left or right) and whether the treasure arrived, and where it arrived. Those that have a non-zero immediate reward and/or a future value give:

$$\begin{aligned}
 Q[s_{13}, a_2] = & \\
 & 0.8 * 0.8 * (0 + 0.9 * 2) \quad \text{— up, no treasure} \\
 + & 0.8 * 0.2 * 0.25 * (0 + 0.9 * 7) \quad \text{— up, treasure at top right} \\
 + & 0.1 * 0.2 * 0.25(10 + 0.9 * 0) \quad \text{— right, treasure appears there}
 \end{aligned}$$

every other value is 0. Note that $0.2 * 0.25$ is the probability that a treasure appears at the top right state.

10. In learning under uncertainty, when the the EM algorithm used? What is the E-step? What is the M-step?

Solution EM algorithm is used for learning probabilities when the value for some variable is not observed in the data. (E.g., the class variable may not be observed). In the E-step, the data is filled in based on the probabilistic model (we get the expected number in the data). In the M-step the probabilities are updated based on the augmented data (we get the maximum likelihood probabilities).

11. Why don't we use the empirical frequencies when learning probabilities from data? That is, if we observe n occurrences of A out of m cases, why shouldn't we just use n/m as our probability estimate?

Solution This gives not very good estimates if m is small or if $n = 0$ or $n = m$. Just because you have not observed something does not mean that it should have probability zero (which means it is impossible).

You should also expect some questions about what you learned from doing your assignment (e.g., about designing features).