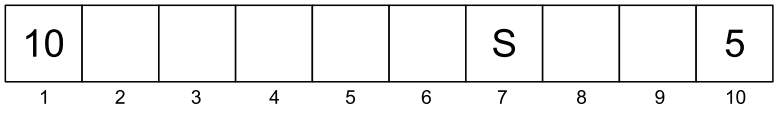
Artificial Intelligence (CS 436)

Homework #4

All work must be individual effort, and use of the web is prohibited (unless otherwise indicated). Any external sources must be cited. Late papers will not be accepted. See the course website for the specific due date.

1. Consider a robot in the following one-dimensional grid-world:



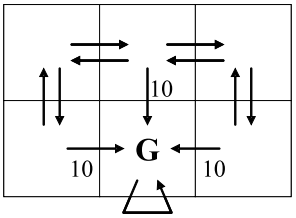
   To find the optimal policy for our robot, we formulate a Markov decision process. Each square corresponds to a state, and the actions from each state are LEFT and RIGHT, which deterministically move the robot one square in the specified direction. "S" corresponds to the robot's start state. The reward is zero in all states except the left and right endpoints, which give +10 and +5 respectively. The MDP has a discount factor of $0 \leq \gamma < 1$. Assume that the endpoints lead immediately to a zero-cost absorbing state. (I.e., these are terminal states, and there are no more rewards after the robot reaches either end and receives either +10 or +5 reward).

   (a) [5 points] If γ is large (i.e., close to 1), then what is the optimal policy while in the start state? What if γ is small (i.e., close to 0)?

   (b) [5 points] Apply value iteration on this problem. Use $\gamma = 0.9$. Fill in the estimated value function for each state in a table until convergence to the optimal policy. Then explain what the optimal policy is.

   (c) [5 points] Repeat the calculation under the assumption movement is in the intended direction only 90% of the time. Otherwise, movement is in the opposite direction. Also, change the discount factor to $\gamma = 0.75$. You only need to show convergence to two decimal places. Is the policy the same as in (b)?

2. Consider the deterministic grid world shown below where the absorbing (i.e., goal) state is denoted **G**. Assume the immediate reward for the labeled transitions is 10 and for the unlabeled transitions is 0.



   (a) [5 points] You are going to apply Q-learning to the grid world assuming the Q table is initialized to all zeroes. Assume the agent begins in the bottom left grid square and then travels clockwise around the perimeter of the grid until it reaches the absorbing goal state, completing the first training episode. Show which Q values are modified as a result of this episode and give their revised values. Answer the question again assuming the agent now performs a second training episode. Finally, answer it again for a third training episode.

(b) [10 points] Once again assuming $\gamma = 0.8$, fill out the entire table, giving the optimal V value for every state in the grid world and the $Q(s, a)$ value for every transition.

(c) [5 points] Suggest a change to the reward function $r(s, a)$ that alters the $Q(s, a)$ values but does not alter the optimal strategy. Then suggest a change to $r(s, a)$ that alters $Q(s, a)$ but does not alter $V(s)$. If you can do both with the same change, that is fine.

3. [10 points] R&N 16.2. Tickets to a lottery cost \$1. There are two possible prizes: a \$10 payoff with probability 1/50, and a \$1,000,000 payoff with probability 1/2,000,000. What is the expected monetary value of the lottery ticket? When (if ever) is it rational to buy a ticket? Be precise—show an equation involving utilities. You may assume current wealth of $\$k$ and that $U((S_k) = 0$. You may also assume that $U(S_{k+10}) = 10 \times U(S_{k+1})$, but you may not make any assumptions about $U(S_{k+1,000,000})$. Sociological studies show that people with lower income buy a disproportionate number of lottery tickets. Do you think this is because they are worse decision makers or because they have a different utility function?

4. R&N 16.3: In 1738, J. Bernoulli investigated the St. Petersburg paradox, which works as follows. You have the opportunity to play a game in which a fair coin is tossed repeatedly until it comes up heads. If the first heads appears on the nth toss, you win 2^n dollars.

   (a) [5 points] Show that the expected monetary value of this game is infinite.

   (b) [5 points] How much would you, personally, pay to play the game? Justify your choice.

   (c) [10 points] Bernoulli resolved the apparent paradox by suggesting that the utility of money is measured on a logarithmic scale (i.e., $U(S_n) = a \lg n + b$, where $S_n$is the state of having $\$n$). What is the expected utility of the game under this assumption?

   (d) [10 points] What is the maximum amount that it would be rational to pay to play the game, assuming that one's initial wealth is $\$k$.

5. R&N 15.9: In this exercise, we analyze in more detail the persistent-failure model for the pattery sensor in Figure 15.13(a) in the textbook.

   (a) [5 points] Figure 15.13(b) in the textbook stops at $t = 32$. Describe qualitatively what should happen as $t \to \infty$ if the sensor continues to read 0.

   (b) [5 points] Suppose that the external temperature affects the battery sensor in such a way that transient failures (i.e., failures that come and go) become more likely as temperature increases. Show how to augment the DBN structure in Figure 15.13(a), and explain any required changes to the CPTs.

   (c) [5 points] Given the new network structure, can battery readings be used by the robot to infer the current temperature? Explain.

6. [10 points] R&N 15.12: Calculate the most probable path through the HMM in Figure 15.20 in the text for the output sequence $[C_1, C_2, C_3, C_4, C_4, C_6, C_7]$. Also give its probability.