Worksheet Week 21

Problems

Q1. Consider the gridworld in figure 1

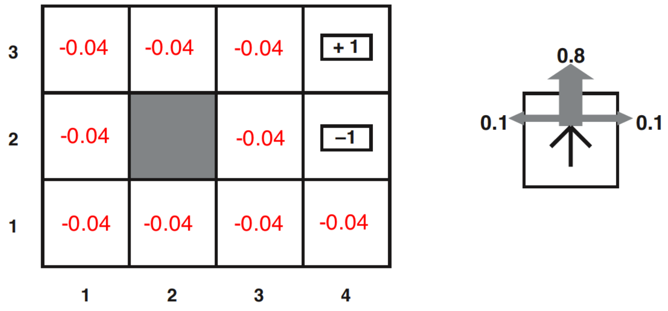

Figure 1: Gridworld

Suppose we wish to use value iteration to compute the utility of each state.

(a) Do we need to wait until the algorithm has converged until we know the utility of each state, or are there some states whose utility we already know?

(b) Suppose we initialise the utility of every state to 0, and then perform one iteration of the value iteration algorithm. What is the utility of each state?

r	-1	10 G
-1	-1	-1
-1	-1	-1

Table 1: 3x3 gridworld

Q2. Consider the 3x3 gridworld shown in table 1. The transition model is as follows: 80% of the time the agent goes in the direction it selects; the rest of the time it moves at right angles to the intended direction, each with a probability of 10%. (i.e., the same as in the previous question).

The r in the top left corner is a reward value. For different values of r, state what policy results. You don't need to use value iteration or policy iteration, you can just work it out from common sense. Use discounted rewards where $\gamma = 0.99$.

(a) $r = -3$

(b) $r = +3$

Q3. Figure 2 shows a narrow bridge represented as a gridworld environment. A robot starts at the left hand side, in the middle row (marked with a reward of 1). The goal is the middle row on the right hand side, marked with a reward of 10. Squares marked with a reward of -100 are terminal nodes, and represent the robot falling off the bridge. The robot can move one square up, down, left or right. When told to move in a specified direction, it moves in the intended direction with probability 0.8 or at 90 degrees to the intended direction with probability 0.1, or at -90 degrees to the intended direction with probability 0.1.

wall	-100	-100	-100	-100	-100	wall
1	0	0	0	0	0	10
wall	-100	-100	-100	-100	-100	wall

wall	-100	-100	-100	-100	-100	wall
1	-17.28 ←	-30.44 ←	-36.56 →	-25.78 →	-10.8 →	10
wall	-100	-100	-100	-100	-100	wall

Figure 2: (a) rewards for the bridge-crossing problem in gridworld. (b) utilities after 5 iterations, and the corresponding optimal policy

(a) Using a discount value of 0.9, calculate the utility of each non-terminal grid square after one and two moves.

(b) The problem leads to the optimal policy shown in Figure 2b, which fails to cross the bridge. What would be the effect on the policy of decreasing the discount value ?

(c) What would be the effect on the policy of increasing the utility of the goal ? Choose a new value for the utility of the goal state so that the optimal policy is to cross the bridge from left to right, and show the utility of each non-terminal grid square after 3 iterations.