Tutorial 3

CZ3005

3.1 Assume that you are a manager of a warehouse (with a maximum capacity of W items). Each month t, you know the current inventory (how many items left) in your warehouse. You might have a guess of the external demand in the next month ($t+1$) with a distribution p (the probability that the external demand are j items is $p(D_t = j)$, $j = 0, 1, 2, \ldots$). Based on this information, you decide to order additional items from a supplier. The cost might come from the storing cost of items in warehouse. Your objective is to maximize the profit. Use your own parameters for fixed costs to buy and store for each item and a fixed selling price.

Please write an MDP formulation for the above problem.
Hint: Decision epochs are made at the beginning of each month, hence all events (more items arrive, fill external orders) would make states change. Actions are the amount of an order.

3.2 This Gridworld MDP operates like to the one we saw in class. The states are grid squares, identified by their row and column number (row first). The agent always starts in state (1,1), marked with the letter S. There are two terminal goal states, (2,3) ✓ with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (North, South, West, or East) happens with probability .8. With probability .1 each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the same state.

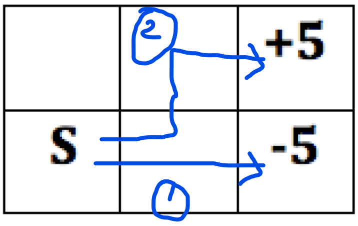

Figure (a) Gridworld MDP.

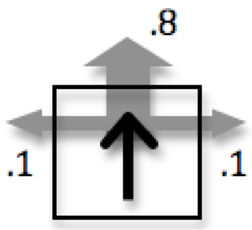

Figure (b) Transition function.

✓ (a) Suppose the agent knows the transition probabilities. Give the first two rounds of value iteration updates for each state, with a discount of 0.9. (Assume V_0 is 0 everywhere and compute V_i for times $i = 1, 2$). (Assume values of termination states ((1, 3) and (2, 3)) are always 0)

✓ (b) Suppose the agent does not know the transition probabilities. What does it need to be able do (or have available) in order to learn the optimal policy?

✓ (c) The agent starts with the policy that always chooses to go right, and executes the following three trials: 1) (1,1) – (1,2) – (1,3), 2) (1,1) – (1,2) – (2,2) – (2,3), and 3) (1,1) – (2,1) – (2,2) – (2,3). What are the Monte Carlo estimates for states (1,1) and (2,2), given these traces (assuming that the discount factor is 1)?

✓ (d) Using a learning rate of .1 and assuming initial values of 0, what updates does the Q-learning agent make after trials 1 and 2, above?